Journal of Machine Learning Research 27 (2026) 1-73 Submitted 1/25; Revised 3/26; Published 5/26

Deep Nonparametric Conditional Independence Tests for Images

Marco Simnacher MARCO.SIMNACHER@HU-BERLIN.DE
Xiangnan Xu XIANGNAN.XU@HU-BERLIN.DE
Chair of Statistics
Humboldt-Universität zu Berlin
Spandauer Str. 1, 10178 Berlin, Germany

Hani Park HANI.PARK@HPI.DE
Christoph Lippert CHRISTOPH.LIPPERT@HPI.DE
Chair Digital Health & Machine Learning
Hasso-Plattner-Institut for Digital Engineering
Prof.-Dr.-Helmert-Straße 2-3, 14482 Potsdam, Germany

Sonja Greven SONJA.GREVEN@HU-BERLIN.DE
Chair of Statistics
Humboldt-Universität zu Berlin
Spandauer Str. 1, 10178 Berlin, Germany

Editor: Brian Kulis

Abstract

Conditional independence tests (CITs) test for conditional dependence between random variables given a vector of conditioning or confounder variables. As existing CITs are limited in their applicability to complex, high-dimensional variables such as images, we introduce deep nonparametric CITs (DNCITs). The DNCITs combine embedding maps, which extract feature representations of high-dimensional variables, with nonparametric CITs applicable to these feature representations. For the embedding maps, we derive general properties on their parameter estimators to obtain valid DNCITs and show that these properties include embedding maps learned through (conditional) unsupervised or transfer learning. For the nonparametric CITs, appropriate tests are selected and adapted to be applicable to feature representations. Through simulations, we investigate the performance of the DNCITs for different embedding maps and nonparametric CITs under varying confounder dimensions and confounder relationships. We apply the DNCITs to brain MRI scans and behavioral traits, given confounders, of healthy individuals from the UK Biobank, confirming null results from a number of ambiguous personality neuroscience studies, now with a larger data set and with our more powerful tests. In addition, in a confounder control study, we apply the DNCITs to brain MRI scans and a confounder set to test for sufficient confounder control. We provide an R package implementing the proposed DNCITs.

Keywords: conditional independence testing, embedding maps, imaging, biomedical data, UK Biobank

1. Introduction

A conditional independence test (CIT) is a statistical test for the null hypothesis of conditional independence, $H_0 : X \perp\!\!\!\perp Y|Z$, between two random variables, X and Y, given a third variable, Z, against the alternative hypothesis of conditional dependence, $H_1 : X \not\perp\!\!\!\perp Y|Z$. X and Y are conditionally dependent, given Z, if X (or Y) provides additional information about Y (or X), given Z. To test the hypothesis, a CIT typically consists of a test statistic that measures the conditional dependence between X and Y, given Z, and a comparison of the observed test statistic with its distribution under the null hypothesis of conditional independence (CI). A CIT should be valid and control the type 1 error (T1E), i.e. the probability of falsely rejecting the null hypothesis, while simultaneously achieving high power, i.e. a high probability of correctly rejecting the null hypothesis of CI.

CITs are used for causal discovery, graphical model learning and feature selection because of the connection between CI and causal inference, prediction sufficiency, and parameter identification (Dawid, 1979). They are currently used in a variety of fields, including biomedical data (Bellot and van der Schaar, 2019; Li et al., 2022; Katsevich and Ramdas, 2022), environmental data (Runge, 2018; Runge et al., 2023), neuroimaging data (Grosse-Wentrup et al., 2016), and economic data (Huang et al., 2016), where data sets are often multimodal and consist of complex objects such as images.

A limitation of existing CITs is their focus on low-dimensional, often univariate X and Y (Li and Fan, 2020). Such CITs cannot easily be applied to complex, high-dimensional objects because they are theoretically or computationally inapplicable, do not hold the T1E in such settings, or lack power against common alternatives. In particular, their test statistics or null hypothesis comparisons are often inapplicable because they are limited to low-dimensional X and Y, either theoretically or computationally. When applicable, CITs often exhibit inflated T1Es due to uncontrolled confounding, as their null hypothesis comparison, designed for low-dimensional X and Y, is inadequate for complex, high-dimensional X or Y with nonlinear relationships to Z. Finally, even when CITs control the T1E, they often lack power as their test statistics are not designed to measure conditional dependence between complex, high-dimensional X and Y given Z. We discuss these limitations in more detail throughout the paper.

Based on these observations, we see three requirements for the proposed deep nonparametric conditional independence tests (DNCITs): (i) Tests must be applicable to complex, high-dimensional random variables, given potential vector-valued confounders. (ii) Tests must control the T1E, even in settings with strong and nonlinear confounding. (iii) Tests should have high power against large sets of alternatives, especially for nonlinear relationships among all random variables combined with small effect sizes. The first requirement is unique to our context, while requirements two and three are standard for (nonparametric) statistical testing, especially nonparametric CI testing (Lehmann et al., 2022; Shah and Peters, 2020). In addition to these requirements, we focus on images X and scalars Y with vector-valued Z throughout this paper due to their wide availability, although the DNCITs can be applied to general complex, high-dimensional X and Y as shown in Appendix B.

To meet these requirements, the proposed DNCITs consist of two steps (Figure 1): First, they map the images to vector-valued feature representations X^ω using embedding maps ω. Second, they apply a suitable nonparametric CIT testing for conditional associations

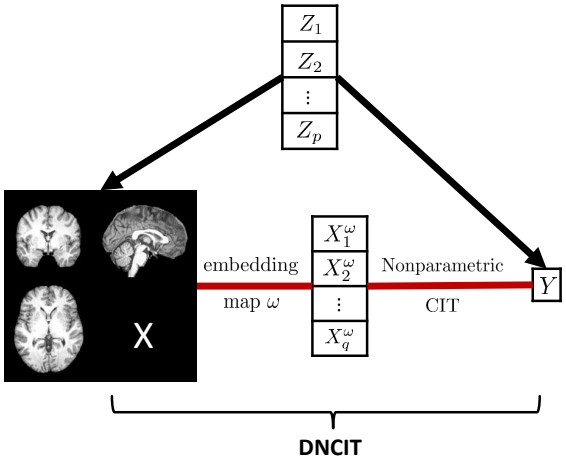

Figure 1: The DNCIT for an image X, a scalar Y and a vector-valued confounder $Z = (Z_1, \ldots, Z_p)$. The black arrows indicate causal effects from Z to X and Y, the red lines represent the two steps of the DNCITs. DNCITs test for conditional dependence between X and Y by mapping X through an embedding map ω to a vector-valued feature representation $X^\omega = (X_1^\omega, \ldots, X_q^\omega)$ in step one and applying a nonparametric CIT which tests for conditional dependence between X^ω and Y given Z in step two.

between X^ω and Y. The dimension reduction of the embedding maps makes the application of existing nonparametric CITs to the embeddings and the scalar feasible. T1E control is ensured through conditionally independently learned embedding maps and the nonlinear confounder control of the nonparametric CITs, as shown in Section 3. The resulting DNCITs have high power to detect nonlinear conditional dependencies between X and Y. This is due to three main advantages of using embeddings. First, embeddings reduce the dimension of the images, mapping the images onto relevant features encoding conditional dependence and reducing noise. Many CITs can only be applied in these lower dimensions. Second, embeddings can simplify the relationship between X^ω and Y and thus simplify detecting the signal of the conditional associations. In particular, for large samples, embedding maps can be trained to encode approximately linear relationships between X^ω and Y under H_1, increasing power of the DNCITs. Third, in particular for small sample sizes, embeddings using transfer learning enable the use of modern, complex image-based models (e.g., MedicalNet) pre-trained on large external data sets, effectively transferring additional data and domain knowledge into the test and thus enhancing the power. Note that our approach is a framework not tied to any particular choice of embedding map and CIT, and could thus also profit from further advances in both areas, while we discuss suitable choices for both theoretically as well as based on simulations in this paper.

In summary, our approach addresses the limitations of existing CITs and provides a unifying framework for CI testing for complex, high-dimensional data such as images. Our contributions include the following:

1. Introduction of DNCITs based on kernels, conditional permutation, conditional mutual information, and prediction models, applicable to an image and a scalar, potentially given a vector-valued confounder.

2. Derivation of theoretical results on the validity of the DNCITs and the T1E excess of one specific DNCIT.

3. Introduction of a simulation design with real-world confounder effects on images X and controllable relationship between X, Z and Y.

4. Evaluation of the T1E and power of DNCITs for different embedding maps and nonparametric CITs, for different confounder dimensions and relationships between the variables, in a simulation study for UK Biobank (UKB) brain MRI scans.

5. Application of DNCITs in two real-world settings: examining the dependence between brain MRI scans and behavioral traits while accounting for confounders, and testing for sufficient confounder control for brain MRI scans in the UKB, extending recent findings in the literature.

6. Implementation of all DNCITs in the publicly available R package DNCIT.

The paper is organized as follows: in Section 2, we review the relevant literature on nonparametric CITs and statistical tests with deep learning (DL) embedding maps. The general framework of DNCITs is introduced in Section 3. We first present theoretical results on the validity of the DNCITs, then derive theoretical conditions for embedding maps that lead to valid DNCITs and study the effect of these embedding maps on the power of the tests. Finally, we discuss nonparametric CITs for application to the feature representations. Simulation results and two applications are presented in Sections 4 and 5, respectively. Section 6 concludes.

2. Review of Relevant Literature

We first discuss current work on nonparametric CITs with respect to their applicability to vector-scalar-valued data, which we build on for the second step of the proposed DNCITs. We then review the literature on combining statistical tests with DL embedding maps, related to the first step of our DNCITs.

2.1 Nonparametric Conditional Independence Tests

The often complex, nonlinear relationships between the image X, Y, and Z can result in complex, nonlinear relationships between the feature representation X^ω, Y, and Z for general embedding maps in the DNCITs. For such complex relationships between X^ω, Y, and Z, nonparametric CITs allow DNCITs to have power while controlling the T1E. However, nonparametric CI testing has been shown to be a hard problem for continuous confounders in the sense that they only have power against certain subsets of alternatives and can only control the T1E for certain subsets of the null hypothesis, without making additional assumptions (Shah and Peters, 2020; Neykov et al., 2021). That said, many nonparametric CITs with different assumptions have been introduced, with a review given

in Li and Fan (2020), and we here discuss and select suitable nonparametric CITs for the DNCIT framework. We categorize nonparametric CITs according to their test statistic and comparison with the null hypothesis. To compare the performance of DNCITs using different CITs, we select CITs from different categories that allow transferability to our setting, i.e., that are applicable to a vector-valued X^ω and a scalar Y, can be computed for potentially continuous vector-valued Z, and are implemented. The included CITs are further discussed in Subsection 3.3 and Appendix D.

2.1.1 Test Statistics

The test statistic is typically a measure of the conditional dependence or deviation from CI, thereby affecting the power and T1E of the DNCITs through the kind of nonlinear relationships between X^ω, Y, and Z detectable in this measure. There are nonparametric CITs with test statistics based on kernels, regression residuals, predictions, ranks, and metrics between distributions.

Kernel-based CITs simplify the problem of CI testing by restricting to kernel embeddings of X, Y and Z in reproducing kernel Hilbert spaces (RKHSs). The test statistics of the Kernel CIT (KCIT; Zhang et al., 2011) and the Randomized conditional Correlation Test (RCoT; Strobl et al., 2019) measure the correlation of the residual functions after mapping X^ω, Y and Z into RKHSs and regressing the respective mappings of X^ω, Y on mappings of Z. The test statistic of Doran et al. (2014) uses the maximum mean discrepancy (MMD) between the observed and conditionally permuted mappings of (X^ω, Y, Z), while Zhang et al. (2023a)'s test statistic is based on an MMD between the joint and the product of the marginal distributions of X^ω and Y. In addition, Huang et al. (2022) propose the kernel partial correlation (KPC) estimated using kernels and geometric graph functionals as an extension of the conditional dependence measure introduced in Azadkia and Chatterjee (2021), which has been studied regarding its power in Shi et al. (2024) and Lin and Han (2023). Often, these kernel-based tests rely on relatively weak assumptions regarding the equivalence of their null hypothesis to the null hypothesis of CI compared to other CITs, resulting in a potentially better T1E control. Additionally, they are often flexible in the dimensions of X^ω, Y, and Z. However, their computation can be slow (Li and Fan, 2020, sec. 3.1.4). Thus, we select the RCoT due to its fast and stable implementation. Furthermore, the KPC is selected because of its stable implementation, and since even non-Euclidean variables would be allowed for X^ω, Y and Z (see Appendix D.1).

Residual-based CITs test for the dependence of the residuals of the regression functions of Y on Z and X^ω on Z. Shah and Peters (2020) propose the Generalized Covariance Measure (GCM), extended in Scheidegger et al. (2022); Lundborg et al. (2024) and Lundborg et al. (2022), which uses a normalized sum of the product of these residuals as the test statistic. In addition to the kernel-based RCoT, which belongs to this category as well, Zhang et al. (2017) also apply kernel-based transformations to X^ω and Y followed by a kernel ridge regression on Z, but using a permutation-based approach for the null comparison. In Duong and Nguyen (2022), the residuals are obtained using conditional normalizing flows for one-dimensional X^ω and Y. These residual-based test statistics mostly rely on the fit of a regression model, which is often a fairly feasible task. However, they often assume low-dimensional X^ω and Y, which can make them computationally or even theoretically

inapplicable to our setting of high-dimensional X^ω. Since the RCoT makes the problem feasible via reducing the dimension of X^ω by mapping it to a lower dimensional feature space and then applying fast linear ridge regression to obtain the residuals, we select it for this category of CITs as well.

Prediction-based CITs test for a significant increase in the accuracy of a prediction of Y using both X^ω and Z compared to using only Z without the information in X^ω. To remove the information in X^ω, the Fast Conditional Independence Test (FCIT; Chalupka et al., 2018) uses decision trees and the Predictive Conditional Independence Test (PCIT; Burkart and Király, 2017) uses an aggregation over multiple prediction models without X^ω, while the CIT based on the Conditional Predictive Impact (CPI; Watson and Wright, 2021) computes the risk over arbitrary losses and Spisak (2022) an R^2 with conditionally independently generated samples $X^{\omega,(m)}$. Similarly, model agnostic prediction-based approaches can also be found in the conditional mean dependence literature (Fisher et al., 2019; Covert et al., 2021; Williamson et al., 2021; Cai et al., 2025; Dai et al., 2022; Lundborg et al., 2024; Williamson et al., 2023). Unlike residual-based approaches, where X^ω appears as output, prediction-based approaches have the advantage of considering the high-dimensional X^ω as predictors in regression models. In addition, they do not require the same dimension for X^ω and Y as many other CITs do. However, the tests often consider only certain parts of the conditional distribution of Y given X^ω and Z such as the mean, instead of the entire distribution, resulting in the weaker conditional mean dependence tests and a potential loss of power. We consider the FCIT and the conditional mean dependence test based on the projected covariance measure (PCM; Lundborg et al., 2024) as representatives of prediction-based CITs, since they are stably implemented.

In contrast, **Metric-based approaches**, which aim to characterize the CI by a metric between the joint and the product of the corresponding marginal distributions of Y and X^ω given Z, consider the entire conditional distribution. The Conditional Distance Independence Test (Wang et al., 2015) achieves this by relying on conditional characteristic functions. Similar metric-based approaches have been proposed in Su and White (2007, 2014); Huang et al. (2016) and Runge (2018). Since they consider the entire distribution of X^ω, Y and Z, a major challenge of these test statistics is the need to estimate some variant of the conditional densities, which suffers from the curse of dimensionality (Li and Fan, 2020). As a representative of metric-based CITs we select the Conditional Mutual Information (CMI) k-nearest-neighbor (knn) (CMIknn) CIT of Runge (2018) due to its flexible, stable implementation, which measures conditional dependence using the conditional mutual information as a test statistic.

Rank-based CITs quantify conditional dependence through concordance in the ranks of Y, and potentially also X^ω, computed within local neighborhoods. Azadkia and Chatterjee (2021) estimate conditional dependence, defined as a nonlinear extension of the R^2, via the normalized difference between the ranks of Y within neighborhoods of X^ω, Z and within neighborhoods of Z alone. For phylogenetic association studies and data with underlying tree dependence structure, Wang et al. (2024) propose to aggregate univariate rank-correlation coefficients from several edges of the tree using a weighted sum or the maximum, and then define local conditional rank correlations by computing these measures within neighborhoods of Z and averaging them to obtain a global statistic. Motivated by conditional copula theory, Ascorbebeitia et al. (2022) propose a Wald-type test based on

a conditional multivariate Kendall's τ, which measures joint rank concordance across all components of Y, X^ω and is therefore mainly suitable for low-dimensional settings. Since we do not consider tree structured data and are in a high-dimensional setting, we include the KPC from above for this category as a kernelized extension of Azadkia and Chatterjee (2021)'s test statistic.

2.1.2 Comparison to the Null Hypothesis

A CIT's T1E control relies on comparing the observed test statistic to its distribution under the null hypothesis. A DNCIT inherits the T1E control from the T1E control of its nonparametric CIT for appropriately chosen embedding maps, see Section 3. There are null comparisons based on (asymptotic) distributions of the test statistic, local permutation, or conditional randomization.

For some test statistics an **(asymptotic) distribution** under the null hypothesis can be derived. The null hypothesis can then be rejected if the test statistic exceeds the corresponding quantile of the distribution. Among others, the CITs of Zhang et al. (2011); Strobl et al. (2019); Shah and Peters (2020) and Scetbon et al. (2022) are based on asymptotic distributions of the corresponding test statistics. This is often preferable in terms of the computational cost of the test, but may suffer from inflated T1Es for small sample sizes if the distribution can only be derived asymptotically. In this category fall the selected RCoT and FCIT.

Local permutation approaches alternatively focus on locally permuting either X^ω or Y or both. Neykov et al. (2021) and Kim et al. (2022) extend local permutation from discrete to continuous confounders by binning Z and permuting Y within these bins. Alternatively, the Classifier CIT (CCIT; Sen et al., 2017) permutes the values of Y for the 1-nearest neighbor in Z, approximately resulting in samples from the null hypothesis. Similarly, the CMIknn (Runge, 2018) randomly permutes Y with one of its knns in Z. While Wang et al. (2024) propose an asymptotic method, the conditional randomization framework described below as well as a knn bootstrap locally permuting Y to compare their test statistic to its distribution under the null, they recommend the knn bootstrap method. Although these approaches are able to obtain approximately conditionally independent samples through simple knn and binning models, they can be computationally expensive and suffer from the curse of dimensionality in the confounder. Due to its computational advantages and potential to trade-off the power and T1E depending on the number of knn (Runge, 2018), the CMIknn is selected for this category.

The **conditional randomization test** (CRT; Barber and Candès, 2015; Candes et al., 2018) requires knowledge on or an approximation of the marginal conditional distributions $\mathbb{P}^{X^\omega|Z}$ or $\mathbb{P}^{Y|Z}$ to compare the observed test statistic with test statistics at generated samples from the marginal conditional distributions. Katsevich and Ramdas (2022) and Wang and Janson (2022) analyse the power of CRT-based statistical tests theoretically. Shi et al. (2021) and Bellot and van der Schaar (2019) propose CRT-based CITs approximating the marginal conditional distributions by Generative Adversarial Networks (GANs). Furthermore, the conditional permutation test (CPT; Berrett et al., 2020), extended in Spisak (2022), was inspired by the CRT. It also requires additional knowledge about $\mathbb{P}^{X^\omega|Z}$ or $\mathbb{P}^{Y|Z}$, but uses a permutation of the observed sample. CRT and CPT provide T1E control,

flexibility in the choice of the test statistic and thus, potentially large power for well chosen test statistics. Due to its superior performance compared to the CRT with respect to the excess T1E (Berrett et al., 2020), the CPT together with the KPC as test statistic is selected.

2.2 Statistical Tests with Embedding Maps

Statistical tests using latent representations typically involve projecting complex, high-dimensional objects onto lower-dimensional feature representations, followed by applying significance tests to these projections. Existing research has primarily focused on either two-sample testing and representation learning or real-world applications of CITs using specific embedding maps.

Specifically, Kirchler et al. (2020) propose two two-sample tests based on the Maximum Mean Discrepancy (MMD) between the feature representations of two high-dimensional objects. Similarly, Liu et al. (2020) enhance the power of a two-sample test by optimizing kernels that generate the feature representations, maximizing test power relative to the learned representations. This approach builds on Sutherland et al. (2021), which optimizes the power of MMD-based two-sample tests over the choice of kernel.

Duong and Nguyen (2022) construct a CIT testing for dependence between the latent representations of conditional normalizing flows for two one-dimensional real-valued variables X, Y, given a d-dimensional real-valued variable Z. Pogodin et al. (2023) derive a conditionally independent feature representation of potentially high-dimensional, complex objects and remark on a potential future extension of this framework to statistical testing.

More directly related to our work, Kirchler et al. (2022) develop a CIT for genome-wide association studies (GWAS), projecting images onto a few leading principal components of feature representations learned using sample splitting or transfer learning, and using these representations in parametric linear genetic association tests. Rakowski et al. (2024) applied this test to UKB data in a large-scale GWAS. Likewise, Kook and Lundborg (2024) applied a conditional mean dependence test based on the projected covariance measure of Lundborg et al. (2024) on three multimodal data sets, including one with pre-trained embeddings of chest x-rays.

While the listed theoretical works combine embedding maps with statistical tests, they either do not consider CITs or do not consider high-dimensional, complex objects (Duong and Nguyen, 2022). Although the above applied studies demonstrate the practical utility of CITs for image-scalar-valued data in specific settings, we propose a general and modular nonparametric CIT framework, provide theoretical requirements for embedding maps to obtain valid CITs, and conduct a comprehensive simulation study exploring the effects of nonparametric CITs and CIT-specifically trained as well as general embedding maps on the T1E rate and power.

3. Deep Nonparametric Conditional Independence Tests

Let $(\mathcal{X}, \mathcal{F}_\mathcal{X}), (\mathcal{Y}, \mathcal{F}_\mathcal{Y}), (\mathcal{Z}, \mathcal{F}_\mathcal{Z})$ be measurable spaces and X, Y, Z be $(\mathcal{X}, \mathcal{F}_\mathcal{X})$-, $(\mathcal{Y}, \mathcal{F}_\mathcal{Y})$-, $(\mathcal{Z}, \mathcal{F}_\mathcal{Z})$-valued random variables with joint distribution $\mathbb{P}^{X,Y,Z}$. Moreover, let $\hat{\beta}$ be a $(\mathcal{B}, \mathcal{F}_\mathcal{B})$-valued random variable representing parameters of $\omega : \mathcal{X} \times \mathcal{B} \to \mathfrak{X}, (X, \hat{\beta}) \mapsto \omega(X, \hat{\beta})$, a $(\mathfrak{X}, \mathcal{F}_\mathfrak{X})$-valued random variable defined through a measurable function ω of

$(X, \hat{\beta})$, called embedding map. We denote the random representation of the random variable X obtained through the embedding map by $X^\omega = \omega(X, \hat{\beta})$. The focus of this paper is on $\mathcal{X}$ as a high-dimensional space of images, $\mathcal{Y} \subseteq \mathbb{R}$ for a univariate continuous variable, $\mathcal{Z} \subseteq \mathbb{R}^p$ for a multivariate confounder (or $p = 0$ for unconditional independence tests), and $\mathfrak{X} \subseteq \mathbb{R}^q$ as the space of feature representations. Moreover, several of the discussed nonparametric CITs are also directly applicable to vector-valued Y, allowing for the corresponding DNCITs to be applied to vector-valued Y as well. For extensions of the framework to complex, high-dimensional Y, such as images, see Appendix B.

We assume that $(X_i, Y_i, Z_i), i = 1, \ldots, n$ are n independently and identically distributed (i.i.d.) copies of (X, Y, Z). Additionally, we define $X_i^\omega = \omega(X_i, \hat{\beta})$, and write $X^n = (X_1, \ldots, X_n)$, $(X, Y)^n = (X_i, Y_i)_{i=1}^n$, and analogously for all other combinations of X_i^ω, X_i, Y_i, Z_i, $i = 1, \ldots, n$. For the samples with images and their feature representations, we write $S = (X, Y, Z)^n$ and $S^\omega = (X^\omega, Y, Z)^n$, respectively. Furthermore, let the underlying probability spaces for the samples be $(\mathcal{S}, \mathcal{F}_\mathcal{S}, \mathbb{P}^S)$ and analogously $(\mathcal{S}^\omega, \mathcal{F}_\mathcal{S}, \mathbb{P}^{S^\omega})$.[1]

The idea of DNCITs is to translate the problem of testing

$$H_0 : X \perp\!\!\!\perp Y | Z \quad vs. \quad H_1 : X \not\perp\!\!\!\perp Y | Z \tag{1}$$

to testing

$$H_0^\omega : X^\omega \perp\!\!\!\perp Y | Z \quad vs. \quad H_1^\omega : X^\omega \not\perp\!\!\!\perp Y | Z \tag{2}$$

with an existing nonparametric CIT. We denote the collection of null and alternative distributions of (1) by $\mathcal{P}_0 = \{\mathbb{P}^S | X \perp\!\!\!\perp Y | Z\}$ and $\mathcal{P}_1 = \{\mathbb{P}^S | X \not\perp\!\!\!\perp Y | Z\}$, respectively, and correspondingly of (2) by $\mathcal{P}_0^\omega = \{\mathbb{P}^{S^\omega} | X^\omega \perp\!\!\!\perp Y | Z\}$ and $\mathcal{P}_1^\omega = \{\mathbb{P}^{S^\omega} | X^\omega \not\perp\!\!\!\perp Y | Z\}$, respectively.

Then, we propose the DNCIT as measurable function

$$\varphi_{n,\omega,\theta} : \mathcal{S} \to \{0, 1\}, S \mapsto \varphi_{n,\theta}^\omega \circ \bar{\omega}(S), \tag{3}$$

where ω is the embedding map depending implicitly on an estimator $\hat{\beta}$ of parameters β, $\bar{\omega}$ takes the whole sample S as input and maps it onto S^ω through the embedding map, the measurable function $\varphi_{n,\theta}^\omega : \mathcal{S}^\omega \to \{0, 1\}, S^\omega \mapsto \varphi_{n,\theta}^\omega(S^\omega)$ is a nonparametric CIT with parameters θ testing (2), and $\varphi_{n,\omega,\theta}(S) = 0$, respectively 1, define decisions for H_0 and H_1, respectively, in (1). In the following, we omit the dependence on n and write $\varphi_{\omega,\theta}$ for the DNCIT and φ_θ^ω for the corresponding nonparametric CIT. The test statistic for a DNCIT is defined as

$$T_{\omega,\theta} : \mathcal{S} \to \mathbb{R}, S \mapsto T_{\omega,\theta}(S) = T_\theta^\omega \circ \bar{\omega}(S), \tag{4}$$

where $T_\theta^\omega : \mathcal{S}^\omega \to \mathbb{R}, S^\omega \mapsto T_\theta^\omega(S^\omega)$ is the test statistic of the nonparametric CIT φ_θ^ω.

There are two components that can be chosen in DNCITs: 1.) the embedding map ω of images onto feature representations depending on an estimator $\hat{\beta}$ of parameters β, 2.) the

1. More precisely, let the underlying probability spaces be $(\mathcal{S}, \mathcal{F}_\mathcal{S}, \mathbb{P}^S) = (\prod_{i=1}^n (\mathcal{X}_i \times \mathcal{Y}_i \times \mathcal{Z}_i), \bigotimes_{i=1}^n (\mathcal{F}_{\mathcal{X}_i} \otimes \mathcal{F}_{\mathcal{Y}_i} \otimes \mathcal{F}_{\mathcal{Z}_i}), \bigotimes_{i=1}^n \mathbb{P}^{(X_i, Y_i, Z_i)}) = ((\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})^n, \mathcal{F}_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}}^{\otimes n}, \mathbb{P}^{(X,Y,Z)^n})$ and analogously $(\mathcal{S}^\omega, \mathcal{F}_\mathcal{S}, \mathbb{P}^{S^\omega}) = ((\mathfrak{X} \times \mathcal{Y} \times \mathcal{Z})^n, \mathcal{F}_{\mathfrak{X}\times\mathcal{Y}\times\mathcal{Z}}^{\otimes n}, \mathbb{P}^{(X^\omega,Y,Z)^n})$.

nonparametric CIT φ_θ^ω testing (2) and consisting of its test statistic T_θ^ω, and the comparison to the null distribution of the test statistic T_θ^ω. To ensure the resulting test is valid, in Subsection 3.1 we first present theoretical results on the relationship between the null hypotheses in (1) and (2), as well as on the validity of a DNCIT given certain assumptions on the corresponding embedding map and nonparametric CIT. Subsection 3.2 translates these theoretical results to the validity of DNCITs using embedding maps obtained from different learning procedures (assuming valid nonparametric CITs), and discusses the effect of these embedding maps on the DNCITs' T1E and power. Finally, potentially suitable nonparametric CITs for these feature representations are studied in Subsection 3.3 and Appendix D.

3.1 Validity of Deep Nonparametric Conditional Independence Tests

In the following, we show that correctness of the null hypothesis for image and scalar in (1) implies that the corresponding hypothesis holds for feature representation and scalar in (2), under an assumption on the embedding map which ensures that it does not introduce conditional dependence. A version of this theorem for non-scalar Y and a corresponding embedding map can be found in Appendix B.

Theorem 1 (Relation of null hypotheses H_0 and H_0^ω) *Let $\hat{\beta}$ be a $(\mathcal{B}, \mathcal{F}_\mathcal{B})$-valued random variable such that $\hat{\beta}$ is conditionally independent of Y^n given $(X, Z)^n$, and let $\omega : \mathcal{X} \times \mathcal{B} \to \mathfrak{X}$ be a measurable function. Then, we have the following:*
If $H_0 : X \perp\!\!\!\perp Y|Z$ holds, then $H_0^\omega : X^\omega \perp\!\!\!\perp Y|Z$ also holds.

All proofs are provided in Appendix A.

The general formulation of the theorem via the estimator $\hat{\beta}$ of the parameters of the embedding map allows us to quantify which information we can use to learn the embedding map in addition to the information within X^n. In particular, this is more general than for example in Dawid (1979), which allows the function ω to be only an in-sample function of X^n. In contrast, the theorem here also allows it to be, for example, an in-sample function of $(X, Z)^n$. We note that the theorem does not establish an equivalence between H_0 and H_0^ω. In particular, for an embedding map with $\hat{\beta}$ satisfying $\hat{\beta} \perp\!\!\!\perp Y^n|(X, Z)^n$, it is possible for H_0^ω to hold while H_0 does not. Thus, the embedding maps learned such that $\hat{\beta} \perp\!\!\!\perp Y^n|(X, Z)^n$ can lead to valid DNCITs, as will be shown in the next theorem. However, the power of the DNCITs can decrease for embedding maps that do not encode the conditional dependence between X and Y given Z in a way detectable for the nonparametric CITs. We translate the theorem into concrete learning approaches for embedding maps and discuss embedding maps further in Subsection 3.2. We also study the DNCITs' performance empirically for several embedding maps in Section 4.

We now use this theorem to obtain a result on the validity of DNCITs for $\mathcal{P}_0$. Remember that a DNCIT is valid for the null hypothesis $\mathcal{P}_0$ given a level $\alpha \in (0, 1)$ and sample size n, if

$$\sup_{\mathbb{P}^S \in \mathcal{P}_0} \mathbb{E}_{\mathbb{P}^S}\left[\mathbb{1}\{\varphi_{\omega,\theta}(S) = 1\}\right] = \sup_{\mathbb{P}^S \in \mathcal{P}_0} \mathbb{P}^S(\varphi_{\omega,\theta}(S) = 1) \leq \alpha.$$

Theorem 2 (Validity of DNCITs) *Let $\hat{\beta}$, ω, and X^ω be as in Theorem 1 and assume that there exists a valid nonparametric CIT φ_θ^ω for $\mathcal{P}_0^\omega$. Then, the DNCIT defined as in* (3) *combining this φ_θ^ω with $\bar{\omega}$ is a valid test for $\mathcal{P}_0$.*

This theorem requires the existence of a valid nonparametric CIT for the feature representation and the scalar. Without additional assumptions, there exists no valid nonparametric CIT that has power against any alternative greater than the level of the test (Shah and Peters, 2020). However, making additional assumptions regarding the knowledge of either $\mathbb{P}^{X|Z}$ or $\mathbb{P}^{Y|Z}$ can result in nonparametric CITs that are valid and have greater power than the level against alternatives (Candes et al., 2018). In the case of DNCITs that assume such additional knowledge, we discuss the availability of this knowledge in our setting in Appendix D. In addition, we derive in Subsection 3.3 the T1E excess of such a DNCIT if an approximation of $\mathbb{P}^{Y|Z}$ can be obtained. For DNCITs based on nonparametric CITs without such additional knowledge, we investigate the corresponding T1E control and power of the DNCITs in our simulations. Furthermore, we consider their null and alternative hypothesis for S^ω and possible effects on the DNCITs' T1E and power in Appendix D.

3.2 Embedding Maps for Deep Nonparametric Conditional Independence Tests

The following corollary provides learning approaches for embedding maps that satisfy the assumptions in Theorems 1 and 2. Subsequently, we discuss these learning approaches and the impact of such embedding maps on the T1E and power.

Corollary 3 *Let $(\mathcal{A}, \mathcal{F}_\mathcal{A})$ be a measurable space.*

a) Let $(\widetilde{X}, \widetilde{Y}, \widetilde{Z})$ be $(\widetilde{\mathcal{X}}, \mathcal{F}_{\widetilde{\mathcal{X}}}), (\widetilde{\mathcal{Y}}, \mathcal{F}_{\widetilde{\mathcal{Y}}}), (\widetilde{\mathcal{Z}}, \mathcal{F}_{\widetilde{\mathcal{Z}}})$-valued random variables and assume that $\widetilde{S} = (\widetilde{X}_i, \widetilde{Y}_i, \widetilde{Z}_i)_{i=n+1}^{n+n'}$ is a sample of i.i.d. copies of $(\widetilde{X}, \widetilde{Y}, \widetilde{Z})$, such that $\widetilde{S}$ is independent of S. In particular, $\widetilde{\mathcal{S}} = (\widetilde{\mathcal{X}}, \widetilde{\mathcal{Y}}, \widetilde{\mathcal{Z}})^n$ can be equal to $\mathcal{S} = (\mathcal{X}, \mathcal{Y}, \mathcal{Z})^n$. Moreover, let $\hat{\alpha} : \widetilde{\mathcal{S}} \to \mathcal{A}$ be a measurable function. Then, $\hat{\alpha}(\widetilde{S}) \perp\!\!\!\perp Y^n | (X, Z)^n$.

b) Let $\hat{\alpha} : (\mathcal{X} \times \mathcal{Z})^n \to \mathcal{A}$ be a measurable function. If $H_0 : X \perp\!\!\!\perp Y | Z$ is true, then $\hat{\alpha}((X, Z)^n) \perp\!\!\!\perp Y^n | (X, Z)^n$.

c) Let $\hat{\alpha} : \mathcal{X}^n \to \mathcal{A}$ be a measurable function. If $H_0 : X \perp\!\!\!\perp Y | Z$ is true, then $\hat{\alpha}(X^n) \perp\!\!\!\perp Y^n | (X, Z)^n$.

In particular, we obtain for a measurable function $\hat{\beta} : \mathcal{A} \to \mathcal{B}$ that $\hat{\beta}(\hat{\alpha}(\widetilde{S})) \perp\!\!\!\perp Y^n | (X, Z)^n$, $\hat{\beta}(\hat{\alpha}((X, Z)^n)) \perp\!\!\!\perp Y^n | (X, Z)^n$ and $\hat{\beta}(\hat{\alpha}(X^n)) \perp\!\!\!\perp Y^n | (X, Z)^n$, respectively.

The corollary allows the use of different learned embedding maps to obtain valid DNCITs. Part a) of the corollary allows, for example, the use of sample splitting and transfer learning. For transfer learning, it is important that the data set used to learn the embedding map and the data set used for the CIT are independent. If the data sets are independent, the embedding map cannot introduce additional information about Y^n into $X^{\omega,n}$, given Z^n—which could otherwise be the case—and thus, under the test distribution, $X^{\omega,n} \perp\!\!\!\perp Y^n | Z^n$ if $X \perp\!\!\!\perp Y | Z$. In particular, this holds irrespective of whether the source and test distributions differ through covariate shift, conditional shift, label/target shift, or combinations

thereof (Zhang et al., 2013), as long as source and test data sets are independent. Such shifts may, however, affect power, and their effect is generally case-specific. Under covariate shift, power may remain stable if the source-trained embedding captures features that are also informative on the support of the test distribution, but it may decrease when the test sample lies in regions poorly represented in the source data. Under label/target shift, power can change through changes in the marginal distribution of Y, and through the induced change in the test-sample distribution of (X, Z). Conditional shifts are potentially more problematic for supervised transfer embeddings, since the conditional relationship learned in the source distribution may differ from the conditional-dependence signal relevant in the test distribution. In all cases, shifts affect power depending on whether the learned embedding maps alternatives $X \not\perp\!\!\!\perp Y \mid Z$ under the test-sample distribution to detectable alternatives $X^\omega \not\perp\!\!\!\perp Y \mid Z$; in the extreme case, an embedding may map such alternatives close to, or even into, H_0^ω.

Parts b) and c) allow, for example, conditional and unconditional unsupervised learning approaches on the same data to be applied to the confounder Z and the image X. To be more concrete, $\hat{\alpha}$ could be the estimator of a (conditional) variational autoencoder (Kingma and Welling, 2013, cVAE) learned on $(X, Z)^n$, and $\hat{\beta}$ could be the estimator of the parameters of the cVAE's encoder. Then ω is chosen to map the image to the vector-valued mean in the latent space of the cVAE. Since the cVAE cannot learn any additional information about Y beyond that contained in X and Z, the corresponding embedding map ω cannot introduce conditional dependence. Thus, for an embedding map ω obtained through a cVAE as described above, $X^\omega \perp\!\!\!\perp Y|Z$ if $X \perp\!\!\!\perp Y|Z$. Once feature representations of images obtained by such embedding maps are available, Theorem 2 ensures the validity of the corresponding DNCITs using valid nonparametric CITs. In contrast, when $\hat{\beta}$ is learned from $g(S)$, where g is some measurable function of S, it does not hold in general that $X^{\omega,n} \perp\!\!\!\perp Y^n|Z^n$ for such embedding maps. Thus, if the valid nonparametric CITs applied to S^ω can detect the corresponding induced conditional dependence, the T1E increases and the DNCITs are no longer valid.

Embedding maps not specific to the DNCIT: Building on this corollary, we focus first on available embedding maps learned by transfer or (conditional) unsupervised learning, and discuss their benefits and drawbacks within the DNCIT framework. Embedding maps that are not learned specifically for the DNCIT have several benefits: First, they eliminate the need to train DNCIT-specific embedding maps. Training these can be challenging, for example for 3D structural brain MRI scans, due to the computational cost, or the small sample sizes, as the median sample size of neuroimaging studies is 23 (Marek et al., 2022). The above results allow the transfer of embedding maps between data sets, such as the embedding map obtained from the Fastsurfer tool (Henschel et al., 2020) for brain MRI scans. Second, even if we train the embedding maps specifically for a DNCIT, we do not need to split the data in the case of transfer or unsupervised learning. This makes the sample size available for the CIT larger, as well as the sample size for learning the embedding. Third, one may be interested in testing multiple hypotheses between the image X and multiple scalars. Then the feature representation X^ω can be used repeatedly without the need to train multiple embedding maps for each test. For example, in genome-wide and brain-wide association studies, the goal may be to find causal genes associated with complex phenotypes represented by the image (Kirchler et al., 2022; Rakowski et al., 2024; Marek et al., 2022).

The feature representations can then be used to test for conditional associations between the image and each gene or gene region, without having to learn a new embedding map for each test. Finally, this is consistent with many current practical approaches to medical images using unsupervised and transfer learning (Valverde et al., 2021; Salehi et al., 2023; Abrol et al., 2021; Raza and Singh, 2021; Marek et al., 2022; Kirchler et al., 2022; Rakowski et al., 2024). Thus, our results provide additional theoretical guarantees and empirical evaluations for these current practices.

Besides these benefits, we discuss two potential pitfalls of not DNCIT-specifically learned embedding maps and possible solutions. First, current practical approaches often apply a Wald test for the significance of the learned feature representations in a linear model with Y as response and X^ω, Z as covariates. However, as the confounding effects for X^ω, Y, Z with complex, high-dimensional X and general embedding maps ω can be highly nonlinear, this can lead to inflated T1Es for the (linear) Wald test.

A second pitfall, generally relevant to testing with embedding maps, is the dependence of the power on the corresponding embedding maps. The power of valid CITs for (2) increases if elements in $\mathcal{P}_1$ are mapped into regions of $\mathcal{P}_1^\omega$ where the corresponding CIT φ_θ^ω has high power. There are a few approaches, however only for two-sample tests, that maximize power over embedding maps given the asymptotic distribution of the test statistic of the feature representations under the alternative hypothesis H_1^ω (Liu et al., 2020), or via sample splitting or transfer learning explicitly learning an embedding map (Kirchler et al., 2020). For CI testing, such optimal embeddings are not available. Nevertheless, embedding maps that do not explicitly optimize the power can lead to a loss in power of the DNCIT if the deviation from CI cannot be detected by the CIT φ_θ^ω. Similar to the first point above, this can be caused in particular by relationships between X^ω and Y that are strongly nonlinear.

Both of these pitfalls can be addressed by using nonparametric CITs for S^ω, controlling for nonlinear effects of Z while testing against nonlinear alternatives, with their choice further discussed in Subsection 3.3, Section 4, and Appendix D.

DNCIT-specific embedding maps: We propose these as an alternative to address the second pitfall of power loss, when general pre-trained embedding maps do not capture the conditional dependence of X and Y given Z in a way detectable by the nonparametric CIT. When sufficient data and compute are available, embedding maps obtained with supervised learning can preserve more signal on the dependence and thus increase power.

Specifically, we train or fine-tune a model (e.g., a ResNet; He et al., 2016) to predict Y from X. Then, we take the penultimate layer as the embedding X^ω. To avoid inflated T1Es, we train the model using either a transfer data set or an internal sample split such that the parameters of this model satisfy Theorem 1 and Corollary 3. Such DNCIT-specific embedding maps have three advantages. First, a DNCIT using such an embedding map learned on independent data inherits the validity from the nonparametric CIT as shown in Theorem 2. Second, given a large enough training sample, such embedding maps are less likely to result in powerless DNCITs because they lead to approximately linear conditional dependence of Y and X^ω under H_1. Third, this approach leverages well-implemented, standard architectures and training pipelines.

Embedding selection: These options can result in several candidate embedding maps $\omega_1, \ldots, \omega_E$. To select an embedding map that is likely to yield high power, we propose a simple, data-driven, validation-split embedding selection criterion. We split the sample

into (i) a validation split used only for embedding map selection and (ii) an independent test split on which the final DNCIT is conducted. For the test split, the final p-value is computed on the whole split for available embedding maps, and it is split again to train DNCIT-specific embedding maps on one split and compute their final DNCIT p-value on the other split. On the validation split, we apply the same nonparametric CIT to each embedding and obtain p-values for $H_0^{\omega_e} : X^{\omega_e} \perp\!\!\!\perp Y|Z,\ e = 1, \dots, E$; we then select the embedding map with the smallest validation p-value. Because the final DNCIT is performed on the independent test split, this selection step does not affect T1E control of the DNCIT. Since we only use one split of the test sample to calculate the final p-value for DNCITs with DNCIT-specific embedding maps, we also only test $H_0^{\omega_e}$ for these DNCITs on a split of the validation set.

To increase the test sample size, and thus the DNCIT's power, we recommend using a larger test split than validation split. Our embedding selection criterion can prevent choosing an embedding map with low power when at least one candidate embedding map is informative on the validation split. However, we note that the embedding selection criterion cannot distinguish the null hypothesis from the case where all candidate embeddings are uninformative. We empirically study this criterion in Subsection 4.2.2.

3.3 Nonparametric Conditional Independence Tests for Vector-Scalar-Valued Data

We apply nonparametric CITs to S^ω to address potential pitfalls in T1E control and power of DNCITs as discussed above. In particular, the confounder control of nonparametric CITs typically accounts for strong, nonlinear confounding, thus leading to T1E control over large sets in $\mathcal{P}_0^\omega$ and $\mathcal{P}_0$ for the corresponding DNCITs. In addition, the nonparametric CITs can detect complex conditional dependencies between the feature representations and the scalar in $\mathcal{P}_1^\omega$, thus increasing the power of the corresponding DNCITs. However, there is no universal best nonparametric CIT, since nonparametric CI testing is a hard problem as discussed above (Shah and Peters, 2020).

Our selection of nonparametric CITs for the DNCIT framework is not exhaustive. Rather, in order to compare the performance of different categories of CITs within the DNCIT framework, CITs from each category are selected. We selected the RCoT, PCM, CPT-KPC, FCIT, and CMIknn, as discussed in Subsection 2.1. For completeness, we discuss in Appendix D for each selected CIT its test statistic and null hypothesis comparison when testing (2), assuming a given fixed embedding map ω, to explore their hypothesis restrictions compared to (2), hyperparameter dependencies, time complexity, and the resulting potential advantages and disadvantages of the tests when applied to vector-scalar-valued data, for which the tests are usually not explicitly designed. Here, we describe the CPT-based CITs (Berrett et al., 2020) as an example of nonparametric CITs, and additionally translate the T1E excess result from Berrett et al. (2020, Theorem 4) to CPT-based DNCITs.

CPT-based CITs allow for null comparison of a separately chosen test statistic. The CPT assumes, in addition to the data S^ω, that either the distribution $\mathbb{P}^{X^\omega|Z}$ or $\mathbb{P}^{Y|Z}$ is known or can be approximated well. Assuming knowledge of the latter, the idea is to obtain samples $S^{\omega,m} = (X^\omega, Y^{(m)}, Z)^n,\ m = 1, \dots, M$, from the null hypothesis $\mathcal{P}_0^\omega$ by

(conditionally) permuting Y^n. The distribution of the permutation π for Y^n is given by

$$(Y^{(m)})^n = Y^n_{\pi^{(m)}}, \qquad \mathbb{P}(\pi^{(m)} = \pi | (X,Y,Z)^n) = \frac{p(Y^n_\pi | Z^n)}{\sum_{\pi' \in \mathcal{R}_n} p(Y^n_{\pi'} | Z^n)}$$

where $\mathcal{R}_n$ is the set of permutations of $\{1,\dots,n\}$, $Y^n_\pi = (Y_{\pi(1)},\dots,Y_{\pi(n)})$ for $\pi \in \mathcal{R}_n$, $p(Y^n|Z^n) \coloneqq p(Y_1|Z_1)\cdots p(Y_n|Z_n)$ and $p(Y_i|Z_i)$ is the density of $\mathbb{P}^{Y_i|Z_i}$. Two efficient sampling algorithms for this distribution are derived in Section 4 of Berrett et al. (2020). Then, assuming that we know the true distribution $\mathbb{P}^{Y|Z}$, we obtain a valid CIT for testing (2) with p-value as in Berrett et al. (2020, thm. 1, 3) by

$$p^\omega \coloneqq \frac{1 + \sum_{m=1}^M \mathbb{1}\left\{|T^\omega_\theta(S^{\omega,m})| \geq |T^\omega_\theta(S^\omega)|\right\}}{1+M}. \tag{5}$$

Instead of full knowledge of $\mathbb{P}^{Y|Z}$, we usually assume to have additional observations to estimate $\mathbb{P}^{Y|Z}$ by $\widehat{\mathbb{P}}^{Y|Z}$. Then, Berrett et al. (2020, Theorem 4) assures for a significance level $\alpha \in [0,1]$ an excess T1E of

$$\sup_{\mathbb{P}^{S^\omega} \in \mathcal{P}_0^\omega} \mathbb{P}^{S^\omega}(p^\omega \leq \alpha) - \alpha \leq \mathbb{E}_{\mathbb{P}^{(Y,Z)^n}}\left[d_{TV}\left\{\mathbb{P}^{Y^n|Z^n}, \widehat{\mathbb{P}}^{Y^n|Z^n}\right\}\right] \tag{6}$$

where $d_{TV}(\cdot,\cdot)$ denotes the total variation distance between two distributions, and $\mathbb{P}^{Y^n|Z^n} = \mathbb{P}^{Y_1|Z_1} \times \ldots \times \mathbb{P}^{Y_n|Z_n}$.

For given $\mathbb{P}^{Y|Z}$, Theorem 2 ensures the validity of the CPT-based DNCITs based on the p-values in (5), given embedding maps chosen as in Theorem 1 and Corollary 3. Interestingly, we can also translate the result on the excess T1E in (6) to CPT-based DNCITs:

Theorem 4 (T1E bound of CPT-based DNCITs) *Let $\hat{\beta}$, ω, and X^ω be as in Theorem 1. Additionally, let $\widehat{\mathbb{P}}^{Y^n|Z^n}$ be an estimate of the true conditional distribution $\mathbb{P}^{Y^n|Z^n}$. Assume that $H_0 : X \perp\!\!\!\perp Y|Z$ is true. For a CPT-based DNCIT and any level $\alpha \in [0,1]$, we obtain*

$$\mathbb{E}_{\mathbb{P}^S}\left[\mathbb{1}\{\varphi_{\omega,\theta}((X,Y,Z)^n) = 1\}\right] \leq \alpha + \mathbb{E}_{\mathbb{P}^{(Y,Z)^n}}\left[d_{TV}\left\{\mathbb{P}^{Y^n|Z^n}, \widehat{\mathbb{P}}^{Y^n|Z^n}\right\}\right], \qquad \mathbb{P}^S \in \mathcal{P}_0.$$

This result on the excess T1E explicitly shows that CPT-based DNCITs targeting $\mathbb{P}^{Y|Z}$ simplify the null comparison significantly, since the complexity of X and X^ω does not impact the T1E control, which can be an advantage over other CITs which additionally consider X^ω in the null comparison. In combination with the CPT, the remaining choice is that of a suitable test statistic, typically a measure of conditional dependence between X^ω and Y given Z, which we discuss together with additional details on the CPT in Appendix D.1. Please note that these results regarding validity and the T1E bound apply only to the CPT-based DNCIT, because they rely solely on the relationship between Y and Z in the null comparison. Therefore, a complex relationship between the embedding X^ω and Z does not lead to an increase in T1E. This is not necessarily the case for the other CITs and their corresponding DNCITs, where T1E control can be affected by a complex relationship between X^ω and Z if not controlled for by the nonparametric CIT used to test

(2). Therefore, we provide the null hypotheses tested, test statistics and null comparisons as well as assumptions for all selected nonparametric CITs in Appendix D. Our simulation study also evaluates the effect of six embedding maps on the T1E control of DNCITs with different nonparametric CITs.

4. Simulation Study

We examine the empirical performance of the DNCITs on simulated data. The DNCIT R package implementing all DNCITs is available at `https://github.com/MSimnach/DNCIT` and the code to reproduce all the results in this and the next section can be found at `https://github.com/MSimnach/dncitPaper`. Throughout our empirical studies we use data from the UKB (Sudlow et al., 2015), and for the images particularly the brain imaging data contained therein. There are many detailed descriptions of the UKB brain imaging data (e.g., Miller et al., 2016), and we rely on standardized brain imaging pipelines (Smith et al., 2024; Zhang et al., 2002; Fischl, 2012), which were also developed and executed on behalf of the UKB (Alfaro-Almagro et al., 2018, 2021; Smith et al., 2022).

4.1 Design

We compare the proposed DNCITs to each other and to a commonly used baseline on brain imaging data from the UKB. We follow the recommendations and terminology for simulation studies in biostatistics in Morris et al. (2019). Further details on specific design choices can be found in Appendix C.1.

Aims: To evaluate the performance of the DNCITs I) for different nonparametric CITs and II) for different embedding maps in realistic data generating mechanisms (DGMs) with respect to A) increasingly complex confounding effects between Z and Y, and B) increasing confounder dimension.

Data generating mechanisms (DGMs): We consider 18 DGMs. For all, we vary the sample size roughly logarithmically over $n = 145, 256, 350, 460, 825, 1100, 1475, 1964, 5000, 10000$. These DGMs extend simulation studies to evaluate CITs for low-dimensional Y and X (Bellot and van der Schaar, 2019; Strobl et al., 2019; Shah and Peters, 2020) to high-dimensional X.

For the images X and confounders Z, we repeatedly resample pairs (X_i, Z_i) with replacement from the UKB to obtain realistic images and correlation structures with the confounders. In the following, individual realisations will be denoted with lowercase letters. As images $\boldsymbol{X} = (\boldsymbol{x}_1^\top, \dots, \boldsymbol{x}_n^\top)^\top$, we resample T1-weighted structural brain MRI scans such that $\boldsymbol{x}_i \in \mathbb{R}^{256\times256\times256}$. Furthermore, the confounders $\mathbf{Z} = (\mathbf{z}_1^\top, \dots, \mathbf{z}_n^\top)^\top$ are resampled as subsets of age, head size, sex, assessment site, assessment date, MRI quality control metric, head location in scanner (4 features), and the first five genetic PCs, s.t. $\mathbf{z}_i \in \mathbb{R}^k, k = 1, 2, 4, 6, 10, 15$ for six different settings with increasing confounder dimension. Here, k denotes the number of confounder variables prior to categorical expansion. All confounders are continuous except sex and assessment site, which are treated as categorical factors; the 23 assessment sites expand into 22 indicators, substantially increasing confounder complexity. These confounders are chosen to make the results relevant to a broader literature, because they are commonly selected confounders for brain imaging data (Hyatt et al., 2020; Alfaro-Almagro et al., 2021), they are among the relevant confounders iden-

tified in the literature related to our behavioral traits and confounder control applications (Avinun et al., 2020; Alfaro-Almagro et al., 2021), and they are commonly adjusted for in alternative potential applications such as GWAS (Kirchler et al., 2022; Rakowski et al., 2024).

Then, we simulate $\mathbf{y} = (y_1, \ldots, y_n)^\top$ from the confounders $\mathbf{Z}$ and feature representations $\mathbf{X}^{\widetilde{\omega}} = (\mathbf{x}_1^{\widetilde{\omega},\top}, \ldots, \mathbf{x}_n^{\widetilde{\omega},\top})^\top$, $\mathbf{x}_i^{\widetilde{\omega}} \in \mathbb{R}^{139}$ of $\boldsymbol{X}$, where $\widetilde{\omega}$ is the FAST embedding map (Zhang et al., 2002; Smith et al., 2024) which maps the images onto feature representations of size $q = 139$. In particular, $\mathbf{y}$ is simulated for $c = 0$ (CI) and $c = 1$ (no CI) as

$$
\begin{aligned}
y_i &= c\mathbf{x}_i^{\widetilde{\omega},\top}\mathbf{w}_x + g_z(\mathbf{z}_i)\mathbf{w}_z + \varepsilon_i, \quad \varepsilon_i \sim N(0,1) \\
&\text{a) linear: } g_z(\mathbf{s}) = \mathbf{s}^\top, \qquad \text{b) squared: } g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j\in\mathcal{J}_c}), \\
&\text{c) complex: } g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2, s_j s_{sex})_{j\in\mathcal{J}_c}, s_{date}^3, s_{date}^4) \\
\mathbf{w}_x &= (w_{x,1}, \ldots, w_{x,q}),\ w_{x,j} = a_{x,j}\delta_{x,j},\ \delta_x \sim N(\mathbf{0}, \mathbf{I}_q),\ a_{x,j} \sim Ber(0.2) \\
\mathbf{w}_z &= (w_{z,1}, \ldots, w_{z,p_z}),\ w_{z,j} \sim N(0,1)
\end{aligned}
\tag{7}
$$

where $p_z = \dim(g_z(\mathbf{s}))$, $\mathcal{J}_c$ denotes the index set of continuous variables, and per seed we standardize the columns of $\mathbf{X}^{\widetilde{\omega}}, \mathbf{Z}$, as well as their linear predictors within the data set. The standardizations balance the effect sizes on $\mathbf{y}$ within $\mathbf{X}^{\widetilde{\omega}}$ and between $\mathbf{X}^{\widetilde{\omega}}$ and $\mathbf{Z}$ across the DGMs.

To address aim A), we increase the complexity of the confounding effects of Z on Y via the functions g_z for the six confounders age, head size, sex, assessment site, assessment date, and MRI quality control metric. For the confounding effects $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j\in\mathcal{J}_c})$, we increase the confounder dimension over $1, 2, 4, 6, 10, 15$ to evaluate aim B). For both variations, we consider CI ($c = 0$) and no CI ($c = 1$).

DNCITs: To evaluate aim II), we vary the feature representations $\mathbf{X}^\omega$ over six different embedding maps ω, since $\mathbf{X}^{\widetilde{\omega}}$, which is used to simulate conditional dependence, is typically unavailable in practice. First, we set $\mathbf{X}^\omega = \mathbf{X}^{\widetilde{\omega}}$ to establish an oracle-type baseline and to evaluate nonparametric CITs when applied to feature representations, for which they are originally not designed. Second, we use parts of the alternative Freesurfer embedding map (Fischl, 2012; Alfaro-Almagro et al., 2018) as feature representation $\mathbf{X}^\omega \in \mathbb{R}^{n\times 165}$. Third, the feature representation $\mathbf{X}^\omega \in \mathbb{R}^{n\times 256}$ is obtained from the latent representation of a cVAE trained on the ADNI data set (Mueller et al., 2005) for the work of Rakowski et al. (2024). Fourth, we extract the feature representations $\mathbf{X}^\omega \in \mathbb{R}^{n\times 512}$ from the last layer of the MedicalNet (Chen et al., 2019), which was trained for segmentation on eight multi-organ data sets, including brain MRI scans (Menze et al., 2014) independent of the UKB, and in particular for transfer learning applications.

Additionally, we evaluate two DNCIT-specific embedding maps. Fitting two models predicting Y from X for each combination of seed (200), DGM (18), and sample size (10) would be computationally infeasible. Thus, we restricted this evaluation to the sample sizes of $n = 256, 460, 1100, 5000$ and 100 seeds over two DGMs, namely, CI and no CI, with age as the confounder ($k = 1$) and $\varepsilon_i \sim N(0, 0.5^2)$. We also study the effects of signal-to-noise-ratio by evaluating the DNCIT-specific embedding maps for $\varepsilon_i \sim N(0, 0.1^2)$ and $\varepsilon_i \sim N(0,1)$ under no CI. For the first DNCIT-specific embedding map, we trained a ResNet-18 model from scratch using one half of the sample (cf. Subsection 3.1). Then, we extracted the

penultimate layer of the trained model as feature representation $X^\omega \in \mathbb{R}^{512}$ for the CIT on the other half of the sample. For the second DNCIT-specific embedding map, we used the pretrained MedicalNet backbone, to which we added a fully connected regression head to predict Y. We fine-tuned this model using half of the sample and extracted the last layer of the backbone as feature representation $X^\omega \in \mathbb{R}^{512}$ for the CIT applied on the other half.

All four non-DNCIT-specific embedding maps correspond to transfer and unsupervised learning approaches (Corollary 3, parts a) and b)), while the two DNCIT-specific embedding maps correspond to a sample splitting approach (part a). Additional details on all embedding maps can be found in Appendix C.1.

In addition, for aim I) we vary the nonparametric CITs as discussed in Subsection 2.1 and 3.3 with implementation details given in Appendix D. Thus, we apply the DNCITs `Deep-RCoT`, `Deep-PCM`, `Deep-CPT-KPC`, `Deep-CMIknn` and `Deep-FCIT`, where for a given embedding map, in the name we replace `Deep` with the corresponding embedding map used in the DNCIT. Additionally, the DNCIT `Deep-Wald` with a parametric Wald test for the significance of all features in the feature representation is used as a baseline.

Target and performance measures: In our simulation study, we evaluate all DNCITs w.r.t. their performance testing the hypothesis in (1). Performance is measured in terms of the T1E, power and runtime. T1E and power are estimated from the rejection rates

$$\widehat{RR} = \frac{1}{n_{\text{sim}}} \sum_{l=1}^{n_{\text{sim}}} \mathbb{1}\{p_l \leq \alpha\}$$

under the null ($c = 0$) and alternative ($c = 1$) hypothesis, respectively, for a significance level of $\alpha = 0.05$, over $l = 1, \ldots, n_{\text{sim}}$ random data generations and using p-values p_l for each DGM. We choose $n_{\text{sim}} = 200$ for computational reasons, resulting at level $\alpha = 0.05$ in a Monte Carlo (MC) standard error (Morris et al., 2019, sec. 5.3) of

$$\text{MC SE}(\widehat{RR}) = \sqrt{\frac{\widehat{RR}(1-\widehat{RR})}{n_{\text{sim}}}} \approx \sqrt{\frac{0.05*0.95}{200}} \approx 0.015.$$

For the 100 seeds used for the DNCIT-specific embeddings, this increases to 0.022. Finally, runtime for the nonparametric CITs was measured on standard CPU-only high-performance computing nodes featuring AMD EPYC 7742 or Intel Xeon 8352Y/8160 processors with AVX2/AVX-512 support. Runtime for the DNCIT-specific embedding maps was measured on a high-performance computing GPU node equipped with an NVIDIA RTX 6000 Ada/Blackwell-class GPU (96 GB memory) and an Intel Xeon 6517 CPU (3.2 GHz). We ignore the runtime to obtain feature representations from available embedding maps, as the embedding maps were already trained and the forward pass for all observations has to be computed only once before all simulations.

4.2 Results

First, note that the DNCITs differ in the sample sizes to which they can be applied in their current implementation. The `Deep-Wald` is not applicable to sample sizes less than or equal to the dimension of the embedding map, i.e. here 256 for the cVAE embedding map and 512 for the MedicalNet and DNCIT-specific embedding maps, plus the dimension

of the confounder. The `Deep-CMIknn` based on a knn search is not practically feasible for sample sizes larger than 1100 due to computational constraints, where we used the runtime of the `Deep-CMIknn` of about 37 minutes per run for sample size 1100 as a cutoff, also shown in Figure 13. For the CMIknn, we based the number of permutations and the number of nearest neighbors on preliminary experiments (see Appendix D.5 and Runge, 2018). Finally, the `Deep-RCoT, Deep-PCM, Deep-FCIT` and `Deep-CPT-KPC` are applicable to all sample sizes. For the `Deep-CPT-KPC`, we learn $\mathbb{P}^{Y|Z}$ in-sample (cf. also details in Appendix D.1). In Berrett et al. (2020, sec. 6.1.1), in-sample approximation led to conservative p-values under the null hypothesis. We evaluate the T1E control under in-sample approximation in our setting, which also allows us to compare all DNCITs using the same data without assuming the existence of additional data on Z, Y. Second, as the results for the FAST embedding map closely resemble those of the Freesurfer embedding map, they are only presented for the experiments with DNCIT-specific embedding maps (cf. similarity of FAST and Freesurfer described in Appendix C.1). Nevertheless, this suggests that the DNCITs are not sensitive to slight variations in embedding maps compared to the true embedding map. Third, we focus on DNCITs with the Freesurfer embedding map for T1E, as there was no significant variation in T1E across the different embedding maps. Finally, to enhance clarity, we excluded DNCITs from the power plots for all sample sizes if they had a power less than 0.15 at the largest sample size, and we excluded them for those sample sizes where they had a T1E greater than 0.15. Accordingly, the power plots in Figure 2 do not show the `Deep-CPT-KPC, Deep-FCIT,` and `cVAE-CMIknn` for any sample size, with the exception of the `Freesurfer-CPT-KPC` and `Freesurfer-FCIT` for confounder dimension 1. Additionally, the `Freesurfer-CMIknn, MedicalNet-CMIknn, Freesurfer-PCM, Deep-RCoT` and `Deep-WALD` are excluded for specific sample sizes. We present more detailed results including QQ-plots of the p-values in Appendix C.2.

4.2.1 Non-DNCIT-Specific Embedding Maps

In this subsection, we present results for available embedding maps not specifically learned for the DNCITs. As results for complex and squared confounder relationships and for confounder dimensions 1 and 2, as well as for 4, 6, 10, and 15 are similar, we display only results for confounder dimension 1 and 10 with squared confounder relationship and confounder dimension 6 with linear and complex confounder relationship (cf. Figure 2). For more detailed results, see Figures 7–13. We focus on the results for aim I) comparing the different nonparametric CITs. Overall, the `Deep-PCM` and the `Deep-RCoT` performed best, with the former being preferable for confounder dimension larger than two.

In particular, the `Deep-PCM` (green) controls the T1E within one MC standard error. However, as the confounder dimension increases, its p-values become more conservative and its power decreases slightly. The `Deep-RCoT` (blue) controls the T1E for a one-dimensional confounder, but its T1Es inflate with increasing confounder dimension. For several confounders, it shows strongly inflated T1Es for smaller sample sizes (up to 1100), likely due to its reliance on asymptotic approximations (see also Appendix D.2; for small sample sizes, the KCIT could be used as an alternative). Additionally, it shows inflated T1Es for very large sample sizes and confounder dimension larger than two, possibly due to the fixed confounder kernel embedding dimension across varying confounder dimensions. Adjusting this

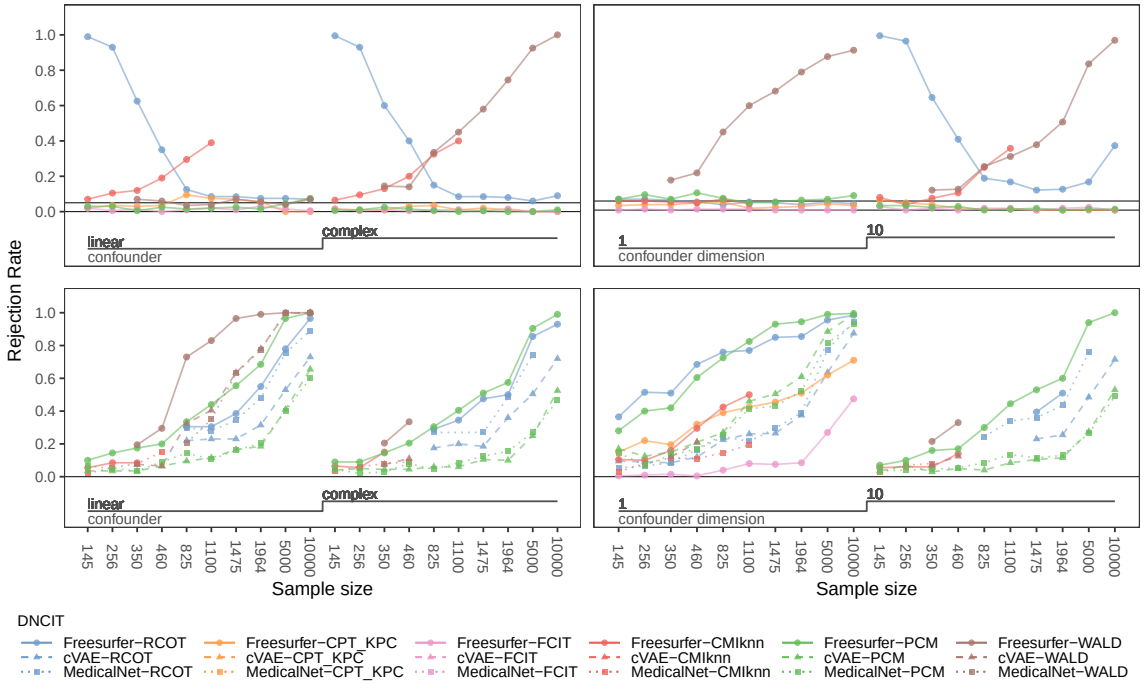

Figure 2: The rejection rates of the DNCITs for $c = 0$ (CI, top row) and $c = 1$ (no CI, bottom row) for increasingly complex confounder relationships (left) and increasing confounder dimension (right). For each column within each panel, the sample size increases from left to right. When the confounder relationship is varied, the confounder dimension is set to 6. When the confounder dimension is varied, the confounder relationship is set to $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j \in \mathcal{J}_c})$, where $\mathcal{J}_c$ denotes the index set of continuous variables. Horizontal lines at 0 and $\alpha = 0.05$.

kernel embedding dimension to scale with the confounder dimension and choosing the penalization parameter of the ridge regressions via cross-validation could improve T1E control, as discussed in Appendix D.2 and Strobl et al. (2019, sec. 7). The `Deep-RCoT` and `Deep-PCM` have comparable power and are the most powerful among all nonparametric tests.

The `Deep-FCIT` (rose), a prediction-based CIT like the `Deep-PCM`, controls the T1E over all settings. However, its power is close to nominal level, except for the `Freesurfer-FCIT` given confounder dimension one and sample sizes 5000 and 10000. The decision tree in the regression model of Y on X, Z does not detect the conditional dependence, demonstrating the importance of well-fitting prediction models for prediction-based DNCITs. The `Deep-CPT-KPC` (orange) maintains T1E across all settings, except for slightly inflated T1Es under linear confounding, which is potentially due to the in-sample approximation of $\mathbb{P}^{Y|Z}$ discussed above. While controlling the T1E, the `Deep-CPT-KPC`'s power is greater than 0.15 only when combined with the Freesurfer embedding map under a one-dimensional confounder. However, this power is lower than that of the `Freesurfer-RCoT` and `Freesurfer-PCM`. The `Deep-CMIknn` (red) only controls the T1E for one- or two-

dimensional confounders. Including categorical sex and assessment site variables resulted in T1Es close to the corresponding power. Reducing the number of local permutations used in the CMIknn could reduce its T1E, but would also reduce its power, see Appendix D.5 and Runge (2018). For one- and two-dimensional confounders, the `Deep-CMIknn` has a medium power, which is comparable to that of the `Freesurfer-CPT-KPC` for one-dimensional confounders. However, this decreases with confounder dimension and other embedding maps.

Finally, the `Deep-WALD`, which only controls for linear confounder effects, has strongly inflated T1Es increasing with sample size in all settings except under linear confounding. In the case of linear confounding, it controls the T1E and has the largest power, followed by the `Deep-PCM` and `Deep-RCoT`.

Figure 22 shows the runtime of the nonparametric CITs. Embedding maps, confounder dimensions and relationships have a minimal effect on runtime. For the `Deep-RCoT`, `Deep-FCIT` and `Deep-WALD`, runtime remains stable as sample size increases. For large sample sizes, the `Deep-RCoT` has the lowest runtime, followed by the `Deep-WALD` and then the `Deep-FCIT`. The runtime of all other DNCITs increases with sample size. The `Deep-PCM` has the lowest runtime for small sample sizes, but this increases to exceed that of the `Deep-FCIT` for larger sample sizes. Finally, the knn-based `Deep-CMIknn` and `Deep-CPT-KPC` have the longest runtime.

The T1E control of the DNCITs is only weakly affected by different embedding maps. In contrast, maximal power is achieved for all DNCITs when using the Freesurfer embedding map, which is closest to the causal FAST embedding used to generate the data. For the other embedding maps, power decreases, although their relative ordering also depends on the underlying nonparametric CIT. We discuss this in more detail in the next subsection, where we also consider DNCIT-specific embedding maps, focusing on the two nonparametric CITs that perform best: the `Deep-RCoT` and `Deep-PCM`.

4.2.2 DNCIT-Specific Embedding Maps, Dimension of the Feature Representation, and Embedding Selection Criterion

First, we focus on the results under noise $\varepsilon_i \sim N(0, 0.5^2)$ in (7). `Deep-RCoT` and `Deep-PCM` both control the T1E and are well calibrated across all embedding maps in the considered DGM with age as confounder (see Figures 14, 15 and 16). In addition, both have similar power (depending also on the embedding) as shown in Figure 3 (and more detailed in Figure 19). Furthermore, the DNCITs with the embedding map trained from scratch clearly outperform those with the fine-tuned MedicalNet. This shows that power also depends on the training process. As before, the power increases with similarity to the FAST embedding map. Thus, the tests perform best with the FAST and Freesurfer embedding maps.

Figure 3 shows that the `Deep-PCM` (solid lines) has the next largest power with the cVAE, followed closely by the MedicalNet and from scratch training (`Scratch-PCM`) for sample sizes smaller than 5000. In contrast, for 5000 observations power is highest for `Scratch-PCM`, followed by `cVAE-PCM` and `MedicalNet-PCM`. The fine-tuned MedicalNet `MedicalNet-ft-PCM` performs worst across all sample sizes. Thus, in our experiments PCM-based DNCITs are most powerful with non-DNCIT-specific embedding maps for typical sample sizes in MRI studies (Marek et al., 2022), and only gain power with DNCIT-specific embedding maps for large sample sizes. This is potentially due to the random forests applied to the feature

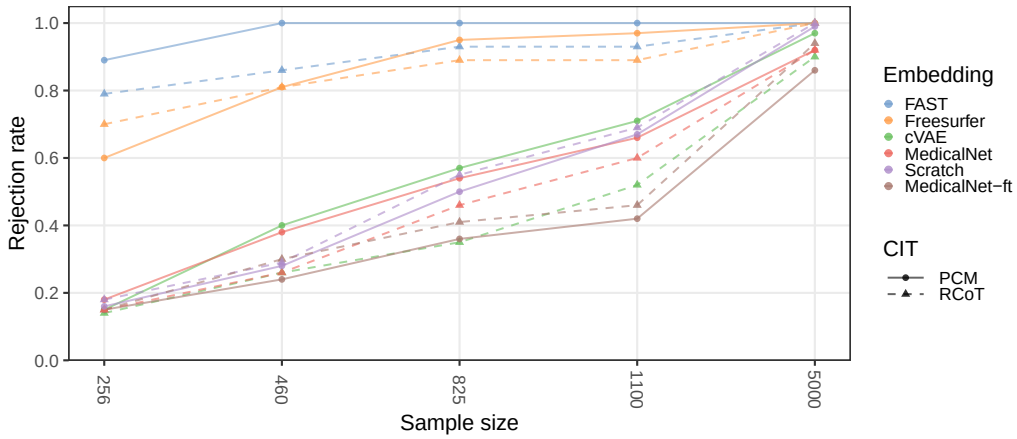

Figure 3: Rejection rates over 100 seeds for DNCITs combining the embedding maps and nonparametric CITs specified in the legend. The sample sizes are equidistantly depicted. The results correspond to the DGM under conditional dependence with $\varepsilon_i \sim N(0, 0.5^2)$, one confounder and the quadratic confounder relationship $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j \in \mathcal{J}_c})$, where $\mathcal{J}_c$ denotes the index set of continuous variables. Higher rejection rates indicate a better performance of the DNCIT, as the DGM assumes conditional dependence.

representation in the PCM, which allow for a data-efficient detection of nonlinear conditional dependencies, cf. Appendix D.4. Thus, even the nonlinear associations between non-DNCIT-specific embeddings and Y can be well detected, and splitting the sample to learn the embedding is not efficient for small samples.

The RCoT-based DNCIT (dashed lines) has a similar or lower power than the PCM-based DNCIT for a given embedding map, though it benefits more strongly from DNCIT-specific embedding maps. In particular, the `Scratch-RCoT` performs as well as the cVAE- and MedicalNet-based RCoT DNCITs for 256 observations and better across all other sample sizes. Furthermore, for 5000 observations, the `Scratch-RCoT` performs similarly to the `FAST-RCoT` and `Freesurfer-RCoT`, and outperforms the `Scratch-PCM` slightly. The `MedicalNet-ft-RCoT` performs similarly to the `cVAE-RCoT` and `MedicalNet-RCoT` for sample sizes up to 1100 and outperforms them for 5000 observations. As described in Appendix D.2, the RCoT relies on approximating kernel ridge regressions of each Y and X^ω on Z by linear ridge regressions. Thus, the linear association between X^ω and Y from DNCIT-specific embedding maps can increase the power by more than the increase seen with the `Deep-PCM` based on random forests.

Next, we evaluate how sensitive the power is to the signal-to-noise ratio in Y, by comparing previous results to those of DGMs with $\varepsilon_i \sim N(0, 0.1^2)$ and $\varepsilon_i \sim N(0, 1)$ depicted in Figures 20 and 21. Due to computational constraints, this analysis was not performed for the fine-tuned MedicalNet, but only for the embedding map trained from scratch, which performed better for $\varepsilon_i \sim N(0, 0.5^2)$. Both `Scratch-PCM` and `Scratch-RCoT` outperform DNCITs with cVAE and MedicalNet for smaller sample sizes when $Var(\varepsilon_i)$ is lower. This

suggests that DNCIT-specific embedding maps can increase power for higher signal-to-noise ratio DGMs at smaller sample sizes than for lower signal-to-noise ratio DGMs.

Regarding runtime, the FAST, Freesurfer, MedicalNet, and cVAE embedding maps have already been trained and thus only require feature representation computation once (across all DGMs) for all images. From scratch training takes longer (a median of 8.5 min for $n = 256$, 4.5 hours for $n = 5000$) than fine-tuning MedicalNet (see Figure 22). This is significantly longer than the average runtime for $n = 5000$ of PCM (ca. 2.5 min) and RCoT (ca. 0.2 min).

Next, we evaluate the impact of the feature representation's dimension. While so far considered embeddings differed in dimension (e.g. 512 for MedicalNet, 256 for the cVAE), differences in embedding types do not allow to disentangle the effect of the dimension alone. We thus separately evaluate the effect of a larger dimension for one specific embedding map, by comparing the flattened penultimate layer to the ultimate layer of the encoder of the cVAE. We run the comparison for fixed squared confounder relationship across all confounder dimensions for the two best performing nonparametric CITs RCoT and PCM. Corresponding results are depicted in Figure 23 for CI and Figure 24 for no CI. The larger feature dimension of the penultimate layer leads to similar T1E control and power as the usual cVAE, although the p-values become more conservative, particularly for the PCM. The runtime is on average longer than when using the cVAE. For 10000 observations, this leads to an increase in the average runtime from 10 to 44 seconds for the RCoT and 3 to 22 minutes for the PCM. These results indicate that the feature dimension is not the main factor for the performance of these two DNCITs, although smaller feature representations can increase power slightly and reduce runtime significantly.

Lastly, we demonstrate the potential increase in the DNCIT's power when selecting the embedding by a selection criterion (cf. Subsection 3.2, and for detailed results, see Figures 25–27). For the same DGM as in Figure 3, we choose the embedding with the lowest p-value of the DNCIT on a validation split. DNCITs with the Freesurfer embedding map have the smallest p-values for almost all seeds on the validation split. This facilitates the selection by the selection criterion, which almost always selects the Freesurfer embedding map. When excluding the Freesurfer embedding map, the embedding map trained from scratch outperforms the others on the DNCIT data set. However, on the smaller validation split used for the selection (with around 1100 observations), the embedding maps produce similar p-values on this validation split. In such cases, the selection by the criterion is less advantageous. Together this indicates that the criterion can help select embedding maps that strongly outperform the others on the smaller validation split, but may require larger validation splits for smaller performance differences.

4.3 Conclusion

The performance of the DNCITs is highly dependent on the chosen CIT, with each CIT responding differently to confounder relationships and dimensions. In this study, the prediction based `Deep-PCM` performed best among the evaluated tests across varying confounder relationships and dimensions. However, as a conditional mean dependence test, the PCM can lose power in settings where only higher moments or distributional shape change (see Appendix D.4 and Lundborg et al., 2024), which was not evaluated here. For one or two con-

founders, `Deep-RCoT` controls the T1E and achieves similar power, while being significantly faster for large sample sizes. However, for more than two confounders, `Deep-RCoT`'s T1Es become inflated, especially for small sample sizes, due to the asymptotic approximations of the RCoT and its sensitivity to the confounder dimension. All other nonparametric CITs clearly perform worse. To improve their performance, the most straightforward adaptation could be replacing the KPC in the `Deep-CPT-KPC`, since this DNCIT effectively controls the T1E, but lacks power in most settings. Finally, the `Deep-Wald`, which is often used to test for conditional associations while controlling for confounders, has strongly inflated T1Es when parametric assumptions are not met, and is thus generally not recommended.

Embedding maps have almost no effect on the T1Es, but can affect the power of the DNCITs. The Freesurfer embedding map was the most powerful here because it is closely related to the FAST embedding map that was used to generate the data under the alternative hypothesis. This highlights the advantage of using embedding maps chosen based on prior knowledge of potential conditionally dependent features in the images. If no embedding map based on prior knowledge is available, the most powerful embedding map depends on the chosen nonparametric CIT. For small sample sizes, most embedding maps achieved similar power, with the `cVAE-PCM` performing best with sample sizes up to 1100. For larger sample sizes, training an embedding map from scratch can increase power significantly, with stronger effects on the RCoT- than on the PCM-based DNCIT. The dimension of the feature representation had only a minimal effect on the tests in our experiments. Finally, the embedding selection criterion was able to select an embedding map that outperformed all others.

5. Real-World Applications

We apply the DNCITs in two applications to evaluate their empirical performance on real-world data sets.

5.1 Brain Structure and Behavioral Traits

This subsection applies DNCITs to study links between brain structure and behavioral traits for healthy individuals in the UKB. Such links are important in the field of personality neuroscience and rely on the idea of the brain as the proximal source of behavior; compare, for example, Yarkoni (2015).

There exist several studies showing associations between behavioral traits such as neuroticism and brain structure (DeYoung et al., 2022; Zhang et al., 2023b). However, these findings could not be replicated in healthy individuals from larger cohorts in systematic replication studies (Kharabian Masouleh et al., 2019; Genon et al., 2022). In particular, little evidence was found for associations between brain structures and the big five behavioral traits (BFBTs; Digman, 1990; Avinun et al., 2020). According to Genon et al. (2022), this ambiguity may arise because finding replicable effects in studies with small sample sizes is often unrealistic due to insufficient power of existing statistical tests. Furthermore, several studies have focused on one-to-one mappings between a specific brain structure and a psychometric measure such as a behavioral trait. This ignores potential interactions between brain structures by relying on linear, often univariate statistical methods, resulting in lower

power for the test on the joint significance of all brain structures and the behavioral traits, in particular when also adjusting for multiple testing.

To address these shortcomings of existing studies, we use the larger cohort of the UKB, increasing the sample size for testing from 1107 in the largest replication study (Avinun et al., 2020) to 8634. Additionally, as shown in the simulation study, DNCITs may have greater power to detect nonlinear, distributional relationships between brain structures and BFBTs compared to linear (Avinun et al., 2020) or machine learning prediction-based CITs similar to the FCIT (Genon et al., 2022). Finally, DNCITs allow us to test for conditional associations between the whole brain MRI scan and each BFBT, thereby maximizing the potential effect sizes by including nonlinear effects and interactions within the whole scan.

In the following, we first replicate the study in Avinun et al. (2020), which found little evidence for associations between certain brain structures derived from brain MRI scans and each BFBT on a cohort of healthy university students obtained from the Duke Neurogenetics Study, on the larger cohort of healthy subjects from the UKB. We then extend the study to test for associations between the whole MRI scans and the BFBTs applying the `Freesurfer-RCoT`, `Freesurfer-PCM`, `FAST-RCoT`, `FAST-PCM`, `Scratch-RCoT`, and `Scratch-PCM`, which in our simulations showed better T1E control than the `Deep-Wald` and high power for sample sizes as large as for the healthy subjects in the UKB.

5.1.1 Brain Morphometry, Behavioral Proxies and Confounders

To obtain a sample of healthy subjects, we follow the preprocessing in Avinun et al. (2020, sec. 2.1). Thus, we excluded participants with diagnoses of cancer, stroke, diabetes requiring insulin, chronic kidney or liver disease, those taking psychotropic or glucocorticoid medications, and those with personality and psychiatric disorders based on the information from Sekimitsu et al. (2022, Supplemental Table 3), van der Meulen et al. (2022), and de Ruijter et al. (2022, Supplemental Table 1).

Using brain morphometric measures from the UKB that are analogous to those in Avinun et al. (2020), we selected 107 measures, including surface area, cortical thickness, subcortical volume, and white matter microstructural integrity, based on a subset of the Freesurfer embedding map. We excluded individuals lacking these brain measures, resulting in a sample size of 8634 to replicate the study by Avinun et al. (2020). For the DNCIT analysis, we used the feature representations derived from the FAST and Scratch embedding maps as a robustness check against the Freesurfer embedding map. After excluding individuals without these feature representations, we obtained a sample of 9115, as the FAST embedding map was available for a slightly larger number of individuals in the UKB.

To obtain BFBT scores, we used a neuroticism score (0-12) from the UKB touchscreen questionnaire during recruitment (Smith et al., 2013). Additionally, we created BFBT proxies for sociability, diligence, and curiosity (0-4), and warmth and nervousness (0-5), similar to those in John and Srivastava (1999). We followed Dahlén et al. (2022); de Ruijter et al. (2022), who used these proxies to study myocardial infarction and stroke risks in the UKB cohort. This results in six BFBT proxies used throughout the analysis.

To include confounders, our study followed the confounding instructions and the conventional set of confounders for the Wald test for brain MRI scans of the UKB (Alfaro-Almagro et al., 2021, sec. 2.9). Specifically, we included age, sex, assessment date, assessment center,

head size, head location measures, and MRI scan quality control discrepancy as confounders. Additionally, squared age, squared assessment date, and an interaction term between age and sex were included for all Wald tests. We also included ethnicity information using the first five genetic PCs of whole-genome SNPs derived in the UKB, similar to Avinun et al. (2020), who used the first four multidimensional scaling coefficients.

5.1.2 Analysis

To replicate Avinun et al. (2020), we used a Wald test to assess the significance of each of the 107 brain morphometric measures when each of the BFBTs was regressed on each measure along with the confounders. Additionally, we used the `Freesurfer-Wald` DNCIT to test for the joint significance of all brain measures when each BFBT was regressed on all brain measures together with the confounders. To compare our new DNCITs, we applied the `Freesurfer-RCoT/PCM` to all brain measures and each BFBT, given the confounders, to capture nonlinear interactions between the brain measures as well as nonlinear relationships between brain measures and each BFBT. A robustness analysis was conducted using the `FAST-RCoT/PCM` to test for conditional associations between the brain MRI scans and each BFBT, as well as collectively across the vector of all BFBTs by applying the `Freesurfer-RCoT` and `FAST-RCoT` to the brain MRI scans, all BFBTs, and the confounders. The latter is directly possible with the `Deep-RCoT`, see Appendix D.2 for details. Finally, as described in Subsection 3.2, we trained the embedding map from scratch to predict neuroticism and applied also the `Scratch-RCoT/PCM` to each BFBT. We selected the RCoT and the PCM as the nonparametric CITs throughout this subsection since they performed best in the simulation study.

5.1.3 Results

We depict the results for individual brain structures and BFBTs in the left panel of Figure 4. The traits are shown on the horizontal axis and for each trait, 107 p-values are obtained, one for each brain measure. The five lowest p-values in Avinun et al. (2020, sec. 3.2) are colored in black. Additionally, all p-values with $-\log_{10}(p) > 2.5$ are colored in red and annotated with their corresponding brain measure. All p-values result in $-\log_{10}(p) < 4.11 \approx \alpha_{bonf}$ where α_{bonf} is the Bonferroni adjusted nominal level.

In the right panel of Figure 4, the results of the `Freesurfer-Wald` applied to all brain measures jointly are shown for each BFBT in the left column. The order of the BFBTs corresponds to the order of the BFBTs in the left panel of Figure 4. In the center-left column, the p-values of the `Freesurfer-RCoT` (points) and `Freesurfer-PCM` (triangles) applied to all brain measures for each trait as well as all BFBTs jointly (larger point, only possible for the RCoT as the PCM is not applicable to multivariate Y) are shown. The center-right column depicts the corresponding p-values for the `FAST-RCoT` (points) and `FAST-PCM` (triangles) and each BFBT, as well as all BFBTs jointly (larger point). Finally, the right column shows the corresponding p-values for the `Scratch-RCoT` (points) and `Scratch-PCM` (triangles) and each BFBT, as well as all BFBTs jointly (larger point).

We highlight three main findings from the analysis. First, after multiple testing adjustment with the Bonferroni method, all p-values of individual tests are above the threshold α_{bonf}, agreeing with the little evidence found for associations between brain morphometry

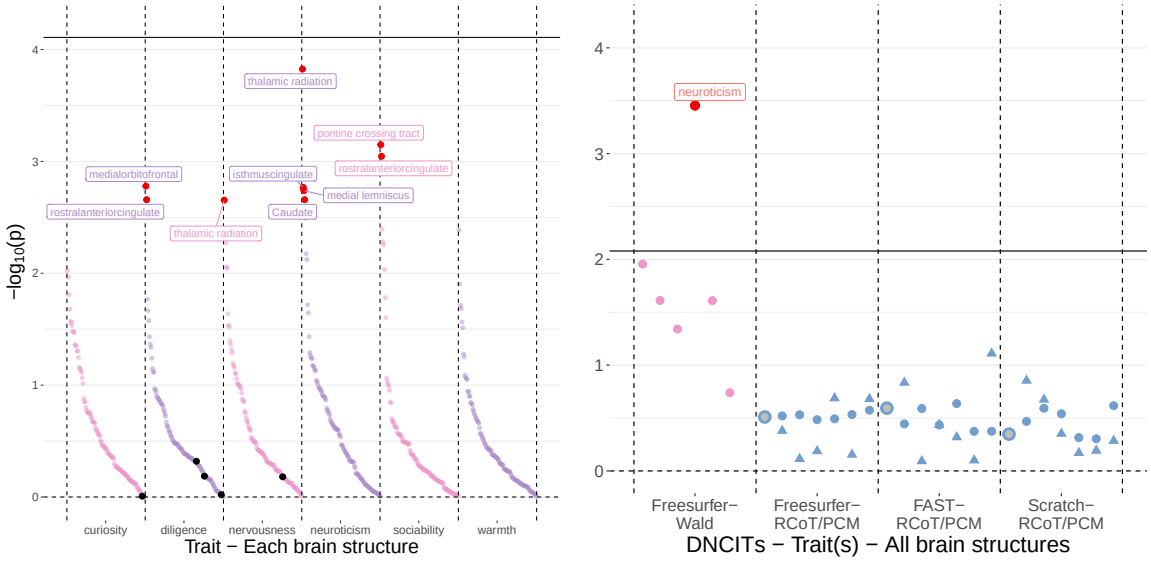

Figure 4: Left panel: $-\log_{10}(p)$ values of individual Wald tests are shown for each trait-brain structure combination, sorted by size for each trait. The trait-brain-structure combinations identified as significant in Avinun et al. (2020) are highlighted in black. Right panel: $-\log_{10}(p)$ values for the `Freesurfer-Wald` (left), `Freesurfer-RCoT/PCM` (center-left), `FAST-RCoT/PCM` (center-right), and `Scratch-RCoT/PCM` (right) across traits with all brain structures. Unicolored points (RCoT) and triangles (PCM) represent curiosity, diligence, nervousness, neuroticism, sociability and warmth (left to right), while grey-filled points represent DNCITs using the RCoT for all BFBTs together. In both figures, tests with $-\log_{10}(p) > 2.5$ are colored red and annotated. The solid vertical lines depict $-\log_{10}(\alpha_{\text{bonf}})$ where α_{bonf} is the nominal level Bonferroni adjusted for 642 (left) and 6 (right) tests for significance level of 0.05, i.e. $\alpha_{\text{bonf}} = \frac{0.05}{642}$ and $\alpha_{\text{bonf}} = \frac{0.05}{6}$, respectively.

measures and BFBT proxies in healthy individuals in Avinun et al. (2020), but now confirmed in the larger cohort of the UKB. The only significant p-value after multiple testing adjustment at a significance level of 0.1 is for the posterior thalamic radiation and neuroticism. Second, for the `Freesurfer-RCoT/PCM`, `FAST-RCoT/PCM`, and `Scratch-RCoT/PCM` we observe that the p-values are larger compared to the `Freesurfer-Wald`. Our simulation study, which indicated strongly inflated T1E for the `Deep-Wald` in case of unaccounted nonlinear confounding, suggests that the smaller p-values of the `Freesurfer-Wald` may only be due to not sufficiently controlling for the nonlinear confounding effects with the confounder set used here. Third, the `Freesurfer-RCoT` and `FAST-RCoT` for the brain imaging data and all behavioral traits are not significant. We expect this test to adequately account for all nonlinear relationships with confounding variables, as shown in the simulation study. Thus, we conclude that there is no evidence for links between the brain structure as measured

by brain MRI scans and personality traits in healthy individuals, also in the larger UKB cohort and using our new more powerful CITs with improved T1E control.

5.2 Confounder Control

The goal of confounder control in hypothesis testing is to ensure that significant associations are not due to the confounders. Thus, the way confounder control is conducted significantly affects the validity of tested associations between psychological traits and brain measures, with inadequate control potentially leading to spurious correlations (Hyatt et al., 2020; Alfaro-Almagro et al., 2021). A common method of confounder control is to linearly regress X and Y on the confounders Z and potentially some pre-defined nonlinear transformations and interaction terms of Z ('regress out'). The resulting residuals are assumed to be controlled for the confounding effects and are used in independence tests. However, there may still be uncontrolled nonlinear and interaction effects that lead to spurious correlations. DNCITs can be used to test whether regressing out does sufficiently control for nonlinear and interaction effects of individual confounders or entire confounder sets in imaging data using Algorithm 1. We illustrate its use on the conventional confounder set for brain imaging data used throughout the previous subsection. The presence of insufficiently controlled effects in the imaging data may result in spurious correlations if the confounder remains associated with X and Y after regressing it out.

Algorithm 1: Test for sufficient control of the effects of confounders on imaging data via regress out

Input: Confounder set $\mathcal{C}$, feature representations X^ω
Output: p-values for uncontrolled confounding effects of all confounders $Z_i \in \mathcal{C}_{base}$

Initialize an empty list *p-values* to store the p-values;
for $Z_i \in \mathcal{C}_{base}$ **do**

1. Linearly regress X^ω on the confounder Z_i and possibly its interactions and nonlinear transformations denoted by $l(Z_i)$ from $\mathcal{C}$; obtain the corresponding residuals $X^{\omega,res}$
2. Test with a nonparametric CIT φ_θ^ω

$$H_0^\omega : X^{\omega,\text{res}} \perp\!\!\!\perp Z_i \mid \mathcal{C} \setminus \{Z_i, l(Z_i)\} \text{ vs } H_1^\omega : X^{\omega,\text{res}} \not\perp\!\!\!\perp Z_i \mid \mathcal{C} \setminus \{Z_i, l(Z_i)\}.$$

3. Store the p-value from the test in *p-values*.

end
return *List p-values.*

To test for sufficient control of the effects of each of the confounders in the conventional confounder set on the brain MRIs, we use the `Residual-RCoT`. This method is similar to the `Deep-RCoT` but involves regressing out a confounder, its pre-specified interaction terms, and nonlinear transformations from the feature representation before applying the RCoT. This allows us to test whether effects from a particular confounder can still be detected, given the remaining conventional confounder set. In particular, we denote the

confounder set consisting of the confounders, their nonlinear and interaction terms as $\mathcal{C}$, and the set consisting of the confounders only as $\mathcal{C}_{base}$. We first regress out one confounder and potentially its nonlinear and interaction terms as defined in the conventional confounder set $\mathcal{C}$ in Subsection 5.1.1 from the brain measures. Then, we use the RCoT to test for conditional associations between the residuals and the confounder from $\mathcal{C}_{base}$, given all other confounders from $\mathcal{C}_{base}$. Applying this algorithm to each confounder in $\mathcal{C}_{base}$, we obtain the following p-values: age 6.7×10^{-2}; sex 3.16×10^{-16}; 5 genetic PCs 2.5×10^{-1}; head size 2.59×10^{-6}; head position 1.9×10^{-2}; assessment date 8.4×10^{-3}; assessment site 2.3×10^{-1}. These results indicate particularly strong effects of sex and head size, even after regressing out these confounders from the brain MRIs and controlling for all other variables in the conventional confounder set. These insufficiently controlled effects in the imaging data may result in spurious correlations if the confounder also remains conditionally associated with the behavioral traits after regressing it out.

This finding is consistent with the results of Alfaro-Almagro et al. (2021), who showed that regressing out the conventional confounder set controls for only part of the variation in brain measures due to confounders. While Alfaro-Almagro et al. (2021) explored the approach of expanding the confounder set by including additional nonlinear and interaction terms, we highlight the possibility of testing for more complex relationships than allowed for in parametric regress out or Wald tests using our DNCITs. Our results in the simulation study and the application in Section 4 and Subsection 5.1 then indicate that the `Deep-RCoT`, the `Deep-PCM` and other DNCITs control for confounding effects more adequately than regressing out the conventional confounder set.

6. Conclusion

We introduced deep nonparametric conditional independence tests (DNCITs) to test for conditional associations between an image and a scalar given vector-valued confounders, providing a theoretically sound framework for CI testing in this context. We presented theoretical results showing the validity of DNCITs using embedding maps, and provided sufficient conditions on the learned embedding map for these to hold, which are fulfilled in addition to sample splitting in particular by transfer and (conditional) unsupervised learning. The latter approaches eliminate the need for sample splitting, leading to potentially more powerful tests and significantly reducing computational costs by utilizing already available embedding maps. Furthermore, we investigated the performance of the DNCITs both theoretically and empirically in a novel simulation design in dependence on the chosen embedding maps and nonparametric CITs. We showed that non-DNCIT-specific embedding maps are computationally efficient and perform better given small samples, especially when informed by domain knowledge, while our proposed DNCIT-specific embedding maps can increase power for larger samples. Moreover, we demonstrated the usefulness of DNCITs in two real-world applications, first in extending recent studies in personality neuroscience and second in testing for sufficient confounder control in brain imaging data. Our empirical results identified nonparametric CITs with both Type I error control and relatively high power, depending on the sample size and confounder dimension. In addition, they highlighted the critical role of confounder control, revealing also the limitations of current practices that rely on parametric tests and their confounder control. Finally, we provide

an R package offering a choice of several nonparametric CITs and useful default embedding maps, particularly for 2D images.

While our focus was on a modular framework with available embedding maps for small sample sizes and DNCIT-specific embedding maps based on robust and standard network architectures for large sample sizes, future work could look at tests that learn optimal embedding maps to maximize power. However, this would require addressing the challenges of post-selection inference, or employing transfer learning or sample splitting, as we did for the DNCIT-specific embedding maps. Furthermore, it would be even more challenging to adapt the standard architectures and loss functions that we relied on for the DNCIT-specific embedding maps to also include the confounder Z theoretically as well as in the implementation in a robust way.

In conclusion, DNCITs enable statistical testing, a key statistical inference tool, for commonly analyzed multimodal data sets of images and tabular data, such as in the biomedical domain. While our focus has been on images, the theoretical framework can also be applied to other multimodal data types, including text.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 459422098. This research has been conducted using the UK Biobank Resource under Application Number 77717. We thank Manuel Pfeuffer, Roshan Rane and Georg Keilbar for helpful discussions. We are grateful to the editorial team and the anonymous reviewers for the valuable comments.

Appendix A. Proofs

Proof [Proof of Theorem 1]For $X \perp\!\!\!\perp Y|Z$, it follows that $X^n \perp\!\!\!\perp Y^n|Z^n$. Additionally, for $\hat{\beta} \perp\!\!\!\perp Y^n|(X,Z)^n$, we have by Dawid (1979, Lemma 4.3) that

$$(X^n, \hat{\beta}) \perp\!\!\!\perp Y^n|Z^n.$$

Since ω is a measurable function of $(X_i, \hat{\beta}), i = 1, \ldots, n$, we apply ω to each $i = 1, \ldots, n$ in $(X^n, \hat{\beta})$ and have by Dawid (1979, Lemma 4.2) for $X^{\omega,n} = (\omega(X_i, \hat{\beta}))_{i=1}^n$ that $X^{\omega,n} \perp\!\!\!\perp Y^n|Z^n$, and thus H_0^ω holds. ■

Proof [Proof of Theorem 2]Let $\mathbb{P}^{S^\omega|\hat{\beta}} = \mathbb{P}^{S|\hat{\beta}} \circ \bar{\omega}^{-1}$ be the conditional probability measure for the measurable map $\bar{\omega}$ of the sample S onto $S^{\omega(\hat{\beta},X)}$. Since we obtain from Theorem 1 that $X^{\omega,n} \perp\!\!\!\perp Y^n|Z^n$, it follows that $\mathbb{P}^{S^\omega|\hat{\beta}} \in \mathcal{P}_0^\omega$. Then,

$$\begin{aligned}
&\sup_{\mathbb{P}^S \in \mathcal{P}_0} \mathbb{E}_{\mathbb{P}^S}\left[\mathbb{1}\{\varphi_{\omega,\theta}(S) = 1\}\right] \\
&= \sup_{\mathbb{P}^S \in \mathcal{P}_0} \mathbb{E}_{\mathbb{P}^{\hat{\beta}}}\left[\mathbb{E}_{\mathbb{P}^{S|\hat{\beta}}}\left[\mathbb{1}\{\varphi_{\omega,\theta}(S) = 1\}\right]\right] \\
&= \sup_{\mathbb{P}^S \in \mathcal{P}_0} \mathbb{E}_{\mathbb{P}^{\hat{\beta}}}\left[\mathbb{E}_{\mathbb{P}^{S|\hat{\beta}}}[\mathbb{1}\{\varphi_\theta^\omega \circ \bar{\omega}(S) = 1\}]\right] \\
&= \sup_{\mathbb{P}^S \in \mathcal{P}_0} \mathbb{E}_{\mathbb{P}^{\hat{\beta}}}\left[\mathbb{E}_{\mathbb{P}^{S|\hat{\beta}} \circ \bar{\omega}^{-1}}[\mathbb{1}\{\varphi_\theta^\omega(S^\omega) = 1\}]\right] \\
&= \sup_{\mathbb{P}^S \in \mathcal{P}_0} \mathbb{E}_{\mathbb{P}^{\hat{\beta}}}\left[\underbrace{\mathbb{E}_{\underbrace{\mathbb{P}^{S^\omega|\hat{\beta}}}_{\in \mathcal{P}_0^\omega}}\left[\mathbb{1}\left\{\varphi_\theta^\omega(S^\omega) = 1\right\}\right]}_{\leq \alpha \text{ for valid } \varphi_\theta^\omega}\right] \leq \alpha
\end{aligned}$$

where we used in line 3 to 4 the integrability of $\varphi_\theta^\omega \circ \bar{\omega}(S)$ under $\mathbb{P}^{S|\hat{\beta}}$ and the resulting equality from Bogachev and Ruas (2007, Thm. 3.6.1).

■

Proof [Proof of Corollary 3]

a) We have $\widetilde{S} \perp\!\!\!\perp S$. Thus, by Dawid (1979, Lemma 4.2(ii)), for the measurable function $U_1 : \mathcal{S} \to (\mathcal{X} \times \mathcal{Z})^n, S \mapsto (X,Z)^n$, we obtain $\widetilde{S} \perp\!\!\!\perp S|(X,Z)^n$. Moreover, by Dawid (1979, Lemma 4.2(i)), for the measurable functions $U_2 : \mathcal{S} \to \mathcal{Y}^n, S \mapsto Y^n$ and $\hat{\alpha}$, it follows:

$$\hat{\alpha}(\widetilde{S}) \perp\!\!\!\perp Y^n|(X,Z)^n.$$

b) Since $X \perp\!\!\!\perp Y|Z$, it follows that $X^n \perp\!\!\!\perp Y^n|Z^n$. Again by Dawid (1979, Lemma 4.1,4.2), we obtain $(X,Z)^n \perp\!\!\!\perp Y^n|(X,Z)^n$. Thus, $\hat{\alpha}((X,Z)^n) \perp\!\!\!\perp Y^n|(X,Z)^n$.

c) Since $X \perp\!\!\!\perp Y|Z$, it follows that $X^n \perp\!\!\!\perp Y^n|Z^n$. Thus, $\hat{\alpha}(X^n) \perp\!\!\!\perp Y^n|(X,Z)^n$.

For a)–c), it follows in particular that $\hat{\beta}(\hat{\alpha}(\cdot)) \perp\!\!\!\perp Y^n|(X,Z)^n$, since $\hat{\beta}$ is a measurable function of $\hat{\alpha}$. ■

Proof [Proof of Theorem 4]For a chosen level α and CPT-based CITs φ_θ^ω, we can rewrite (6) to

$$\sup_{\mathbb{P}^{S^\omega} \in \mathcal{P}_0^\omega} \mathbb{E}_{\mathbb{P}^{S^\omega}} \left[\mathbb{1}\{\varphi_\theta^\omega(S^\omega) = 1\}\right] \leq \alpha + \mathbb{E}_{\mathbb{P}^{(Y,Z)^n}} \left[d_{TV}\left\{\mathbb{P}^{Y^n|Z^n}, \widehat{\mathbb{P}}^{Y^n|Z^n}\right\}\right], \qquad \mathbb{P}^{S^\omega} \in \mathcal{P}_0^\omega.$$

Then, as in the proof of Theorem 2, it follows that

$$\begin{aligned}
\sup_{\mathbb{P}^S \in \mathcal{P}_0} \mathbb{E}_{\mathbb{P}^S} \left[\mathbb{1}\{\varphi_{\omega,\theta}(S) = 1\}\right] &\leq \sup_{\mathbb{P}^S \in \mathcal{P}_0} \mathbb{E}_{\mathbb{P}^{\hat{\beta}}} \left[\sup_{\widetilde{\mathbb{P}} \in \mathcal{P}_0^\omega} \mathbb{E}_{\widetilde{\mathbb{P}}} \left[\mathbb{1}\{\varphi_\theta^\omega(S^\omega) = 1\}\right]\right] \\
&\leq \sup_{\mathbb{P}^S \in \mathcal{P}_0} \mathbb{E}_{\mathbb{P}^{\hat{\beta}}} \left[\alpha + \mathbb{E}_{\mathbb{P}^{(Y,Z)^n}} \left[d_{TV}\left\{\mathbb{P}^{Y^n|Z^n}, \widehat{\mathbb{P}}^{Y^n|Z^n}\right\}\right]\right] \\
&= \alpha + \mathbb{E}_{\mathbb{P}^{(Y,Z)^n}} \left[d_{TV}\left\{\mathbb{P}^{Y^n|Z^n}, \widehat{\mathbb{P}}^{Y^n|Z^n}\right\}\right].
\end{aligned}$$

■

Appendix B. Conditional Independence after Transformations

This section discusses a generalization of Theorem 1 which allows for transformations of X and Y. Therefore, let $(\mathfrak{Y}, \mathcal{F}_\mathfrak{Y}), (\mathcal{E}, \mathcal{F}_\mathcal{E})$ be measurable spaces, $\hat{\gamma}$ be a $(\mathcal{E}, \mathcal{F}_\mathcal{E})$-valued random variable and $\rho : \mathcal{Y} \times \mathcal{E} \to \mathfrak{Y}$ be a measurable function of Y and $\hat{\gamma}$, and denote $Y^\rho = \rho(Y, \hat{\gamma})$. Based on this, we introduce the hypotheses

$$H_0^{\omega,\rho} : X^\omega \perp\!\!\!\perp Y^\rho|Z \quad vs. \quad H_1^{\omega,\rho} : X^\omega \not\perp\!\!\!\perp Y^\rho|Z. \tag{8}$$

Then, $X \perp\!\!\!\perp Y|Z$ implies $X^\omega \perp\!\!\!\perp Y^\rho|Z$ under analogous assumptions on the parameters $\hat{\beta}$ and $\hat{\gamma}$ of the embedding maps ω and ρ as in Theorem 1:

Theorem 5 (Relation of null hypotheses H_0 and $H_0^{\omega,\rho}$ after transformations) *Let $(\mathfrak{X}, \mathcal{F}_\mathfrak{X}), (\mathfrak{Y}, \mathcal{F}_\mathfrak{Y}), (\mathcal{B}, \mathcal{F}_\mathcal{B}), (\mathcal{E}, \mathcal{F}_\mathcal{E})$ be measurable spaces. Furthermore, let $\hat{\beta}$ be a $(\mathcal{B}, \mathcal{F}_\mathcal{B})$-valued random variable such that $\hat{\beta}$ is conditionally independent of Y given X, Z, and let $\omega : \mathcal{X} \times \mathcal{B} \to \mathfrak{X}$ be a measurable function of X and $\hat{\beta}$ and denote $X^\omega = \omega(X, \hat{\beta})$. Moreover, let $\hat{\gamma}$ be a $(\mathcal{E}, \mathcal{F}_\mathcal{E})$-valued random variable such that $\hat{\gamma}$ is conditionally independent of X^ω given Y, Z, and let $\rho : \mathcal{Y} \times \mathcal{E} \to \mathfrak{Y}$ be a measurable function of Y and $\hat{\gamma}$ and denote $Y^\rho = \rho(Y, \hat{\gamma})$.*
If $H_0 : X \perp\!\!\!\perp Y|Z$ is true, then $H_0^{\omega,\rho} : X^\omega \perp\!\!\!\perp Y^\rho|Z$ is true.

Proof By Theorem 1 it follows that $X^\omega \perp\!\!\!\perp Y|Z$. Furthermore, by Dawid (1979, Lemma 4.3), for $X^\omega \perp\!\!\!\perp Y|Z$, $X^\omega \perp\!\!\!\perp \hat{\gamma}|(Y,Z)$, it follows that

$$X^\omega \perp\!\!\!\perp (Y, \hat{\gamma})|Z.$$

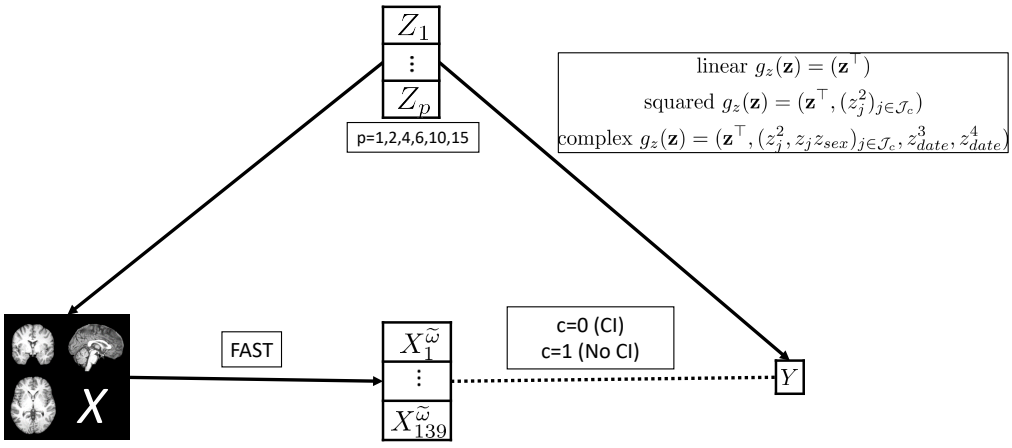

Figure 5: Illustration of the 18 DGMs used in the simulation study. Solid black lines represent the relationships between variables, while the dotted black line indicates either conditional independence ($c = 0$) or dependence ($c = 1$). The feature representation, denoted by $X^{\widetilde{\omega}} = (X_1^{\widetilde{\omega}}, \ldots, X_{139}^{\widetilde{\omega}})$, is derived from the FAST embedding map and is used to simulate the conditional (in)dependence between X and Y. The confounders Z are connected to the brain MRI scans based on their real-world associations in the UKB. We vary the confounder dimension p across 1, 2, 4, 6, 10, and 15 while fixing the confounder relationship to $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j\in\mathcal{J}_c})$. Additionally, the confounder relationship is varied over $g_z(\mathbf{s}) = \mathbf{s}^\top, g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j\in\mathcal{J}_c})$, and $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2, s_j s_{sex})_{j\in\mathcal{J}_c}, s_{date}^3, s_{date}^4)$ while the confounder dimension is fixed to 6. Y is generated according to Equation (7), as a linear combination of $X^{\widetilde{\omega}}$ and Z transformed by one of the functions g_z ($\mathcal{J}_c$ denotes the index set of continuous variables).

Since ρ is a measurable function of $(Y, \hat{\gamma})$, we have by Dawid (1979, Lemma 4.2) that $X^\omega \perp\!\!\!\perp \rho(Y, \hat{\gamma})|Z$, i.e. $X^\omega \perp\!\!\!\perp Y^\rho|Z$. ■

This is more general than formulations such as in Dawid (1979), since we not only allow for functions $\widetilde{\omega} : \mathcal{X} \to \mathfrak{X}$ of X, but instead quantify exactly the additional information that the parameters $\hat{\beta}$ of the embedding maps $\omega : \mathcal{X} \times \mathcal{B} \to \mathfrak{X}$ are allowed to use (analogously for ρ) to lead to conditionally independent feature representations under the null hypothesis (1). In particular, $\hat{\beta}$ can be obtained from all information except that in Y, given X, Z. Thus, we allow the embedding maps to use information in Z and X or Y, respectively, for ω and ρ, respectively. We discussed learning methods for embedding maps that use only such information in and after Corollary 3.

Appendix C. Simulation Study Details

An additional illustrative overview of all simulation designs and DNCITs can be found in Figures 5 and 6.

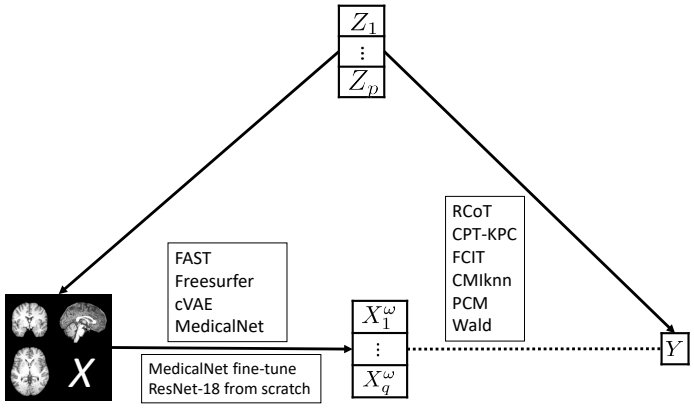

Figure 6: Illustration of the DNCITs applied in the simulation study. Solid black lines represent the relationships between variables, while the dotted black line indicates conditional (in)dependence to be tested. For each (nonparametric) CIT, we apply the embedding maps FAST $(X_1^\omega, \ldots, X_{139}^\omega)$, Freesurfer $(X_1^\omega, \ldots, X_{165}^\omega)$, cVAE $(X_1^\omega, \ldots, X_{256}^\omega)$, and MedicalNet $(X_1^\omega, \ldots, X_{512}^\omega)$. Additionally, we apply two DNCIT-specific embedding maps, fine-tuning the MedicalNet $(X_1^\omega, \ldots, X_{512}^\omega)$ and training from scratch a ResNet-18 $(X_1^\omega, \ldots, X_{512}^\omega)$. The (nonparametric) CITs are varied over the RCoT, PCM, CPT-KPC, CMIknn, FCIT and the Wald test.

C.1 Simulations' Parameter Choices

The sample sizes are chosen to be approximately linear on the logarithmic scale, as in Marek et al. (2022), where the authors used these sample sizes in a study on the reproducibility of significant findings in brain-wide association studies (BWAS). We also include $n = 5000, 10000$ to account for the large number of observations available in the UKB.

Regarding the DGMs: We repeatedly resample pairs (X, Z) from the UKB to obtain samples with the true association of Z and X in the UKB. This makes the results both realistic and particularly relevant to the UKB brain imaging data, but one should be careful about extrapolating the results to other data sets, as the associations between X and Z may differ. In addition, the choice to simulate the conditional dependence between X and Y by the FAST feature representations $\mathbf{X}^{\widetilde{\omega}}$ makes the results particularly relevant to coarse anatomical measures obtained e.g. by the FAST, Fastsurfer or the Freesurfer pipeline (Zhang et al., 2002; Fischl, 2012; Alfaro-Almagro et al., 2018; Henschel et al., 2020), which are often used in BWAS (Hyatt et al., 2020; Marek et al., 2022).

Regarding the embedding maps of the DNCITs: The FAST embedding map (Smith et al., 2024) sums the voxel-wise grey matter partial volume estimates from the FMRIB's Automated Segmentation Tool (Zhang et al., 2002) over 139 atlas-defined parcellations. These estimates are calculated per subject in an unsupervised manner. This provides a coarse tissue-based anatomical representation of the brain MRI scan, consisting of volume measurements of brain structures. The FAST feature representations are used

to model the conditional dependence and serve as a baseline for evaluating nonparametric CITs when the embedding map is known.

The Freesurfer embedding map combines probabilistic atlases of brain structures, parametric MRI intensity models, and rule-based topological corrections. Both the atlases and the intensity models were estimated once from small, manually labeled training sets and are fixed for application to new data. The Freesurfer embedding map closely resembles the true feature representation because it also measures brain structures through volume measurements in regions included in the true representation. However, we did not include all volume measurements of the FAST feature representation. Furthermore, the Freesurfer embedding map includes additional area, volume, surface, and thickness measurements. Thus, it represents a feature set with slightly different and additional information.

The MedicalNet embedding map (Chen et al., 2019; Cardoso et al., 2022) was trained for transfer learning on eight multi-organ segmentation data sets from 3D medical imaging. One of these data sets consists of 3D brain MRI scans (Menze et al., 2014) independent of the UKB. MedicalNet was trained to segment several brain structures, thus providing a potentially suitable, fixed, supervised embedding map that was explicitly proposed for transfer learning.

The cVAE embedding map was trained on ADNI for three tasks (Rakowski et al., 2024): reconstructing the images using a VAE architecture, predicting a clinical dementia rating adding a predictor to the architecture, and learning latent representations invariant to age and sex by conditioning on both variables in the decoder and predictor. The invariance of the latent representations potentially facilitates confounder control for age and sex and increases the power of the DNCITs.

The four feature representations decrease in similarity to the true FAST feature representation. Since the Freesurfer, MedicalNet and cVAE embedding maps are trained on data independent of the UKB, they correspond to transfer learning approaches in Corollary 3.

In addition, we include two DNCIT-specifically learned embedding maps based on sample splitting. First, we fine-tune the ResNet-18 MedicalNet with a simple regression head for the last layer of the MedicalNet backbone, as described by Menze et al. (2014), mapping to Y, using the AdamW optimizer (Loshchilov and Hutter, 2017). The learning rate (lr) of the head is set to 1.2×10^{-3}, the lr of the backbone to 8×10^{-5}, the weight decay to 5×10^{-5}, and a linear warm-up starting at 1×10^{-8} over seven epochs, followed by cosine decay (Loshchilov and Hutter, 2016) for a total of 100 epochs. Furthermore, we unfreeze each layer of the backbone stepwise after eight epochs to increase stability. Second, we train a 3D ResNet-18 architecture from scratch with the AdamW optimizer, an lr of 3×10^{-4}, a weight decay of 1×10^{-7}, and the same scheduler as for fine-tuning. The weights are still initialized using the default initialization for the ResNet-18 in Monai (Cardoso et al., 2022), providing good starting values. For all settings we use a batch size of 16. We rely on PyTorch (Paszke et al., 2019) and the network architectures from Monai (Cardoso et al., 2022) for the implementation of MedicalNet and the model training of the DNCIT-specific embedding maps.

In summary, the simulation study represents settings following findings from large studies on brain imaging data (Hyatt et al., 2020; Alfaro-Almagro et al., 2021; Marek et al., 2022), especially in terms of the confounding effects. Finally, the code and the introduced

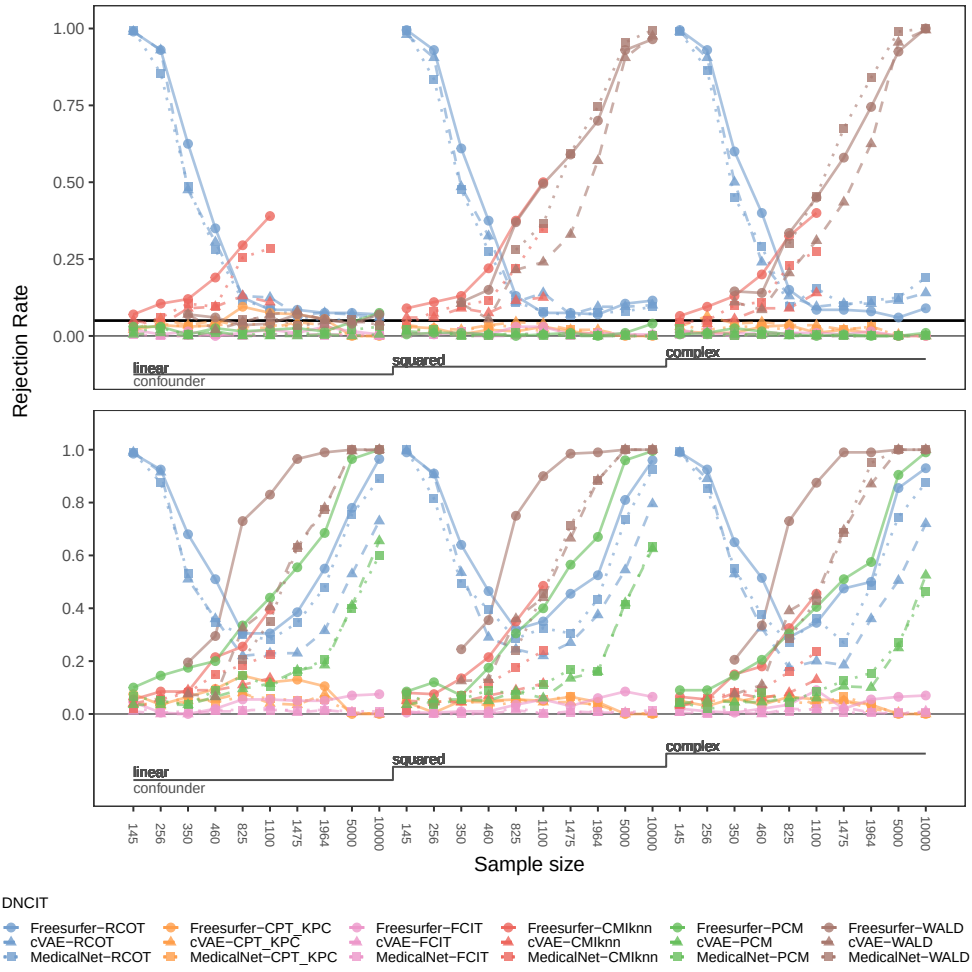

Figure 7: The rejection rates of the DNCITs for $c = 0$ (CI, top) and $c = 1$ (no CI, bottom) for increasingly complex confounder relations (columns). For each column, the sample size is increased from left to right. The confounder dimension is set to 6. Horizontal lines at 0 and $\alpha = 0.05$.

designs can be used to evaluate the performance of the DNCITs on other imaging data sets before applying them to the corresponding real-world application.

C.2 Detailed Results

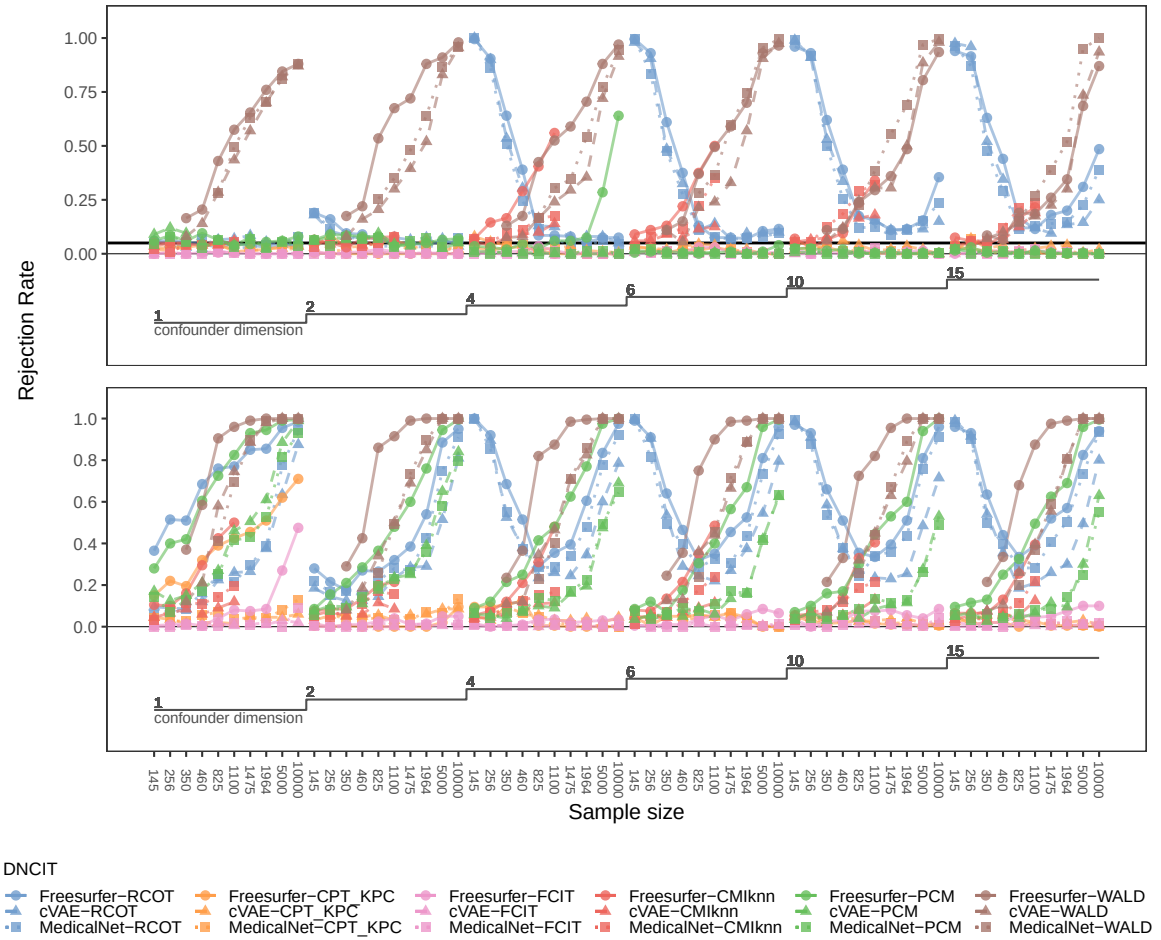

Figure 8: The rejection rates of the DNCITs for $c = 0$ (CI, top) and $c = 1$ (no CI, bottom) for increasing confounder dimension (columns). For each column, the sample size is increased from left to right. The confounder relationship is set to $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j \in \mathcal{J}_c})$, where $\mathcal{J}_c$ denotes the index set of continuous variables. Horizontal lines at 0 and $\alpha = 0.05$.

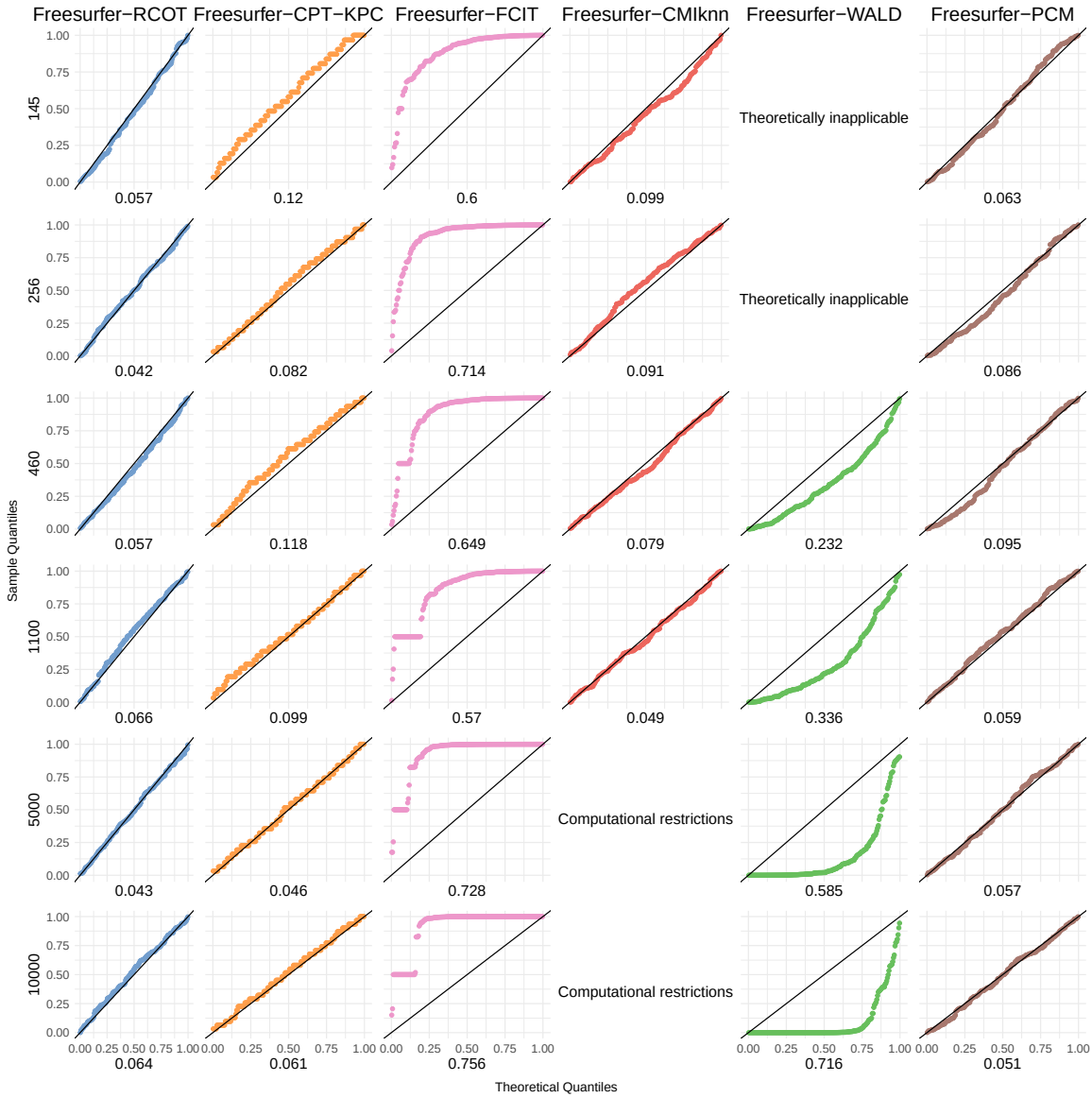

Figure 9: QQ-plots of the observed p-values against the theoretical quantiles of a uniform distribution on the interval $[0, 1]$ for all nonparametric CITs applied to the Freesurfer embedding map at selected sample sizes (indicated at the y-axis label). The results correspond to the DGM under conditional independence with one confounder and the quadratic confounder relationship $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j \in \mathcal{J}_c})$, where $\mathcal{J}_c$ denotes the index set of continuous variables. The diagonal black lines $y = x$ serve as references for the theoretical quantiles of the uniform distribution on $[0, 1]$. If the p-values are uniformly distributed, they should align along this line. The x-axis labels display the Kolmogorov-Smirnov test statistics used to assess whether the p-values are uniformly distributed; smaller values indicate a better calibration of the CIT, as the DGM assumes conditional independence.

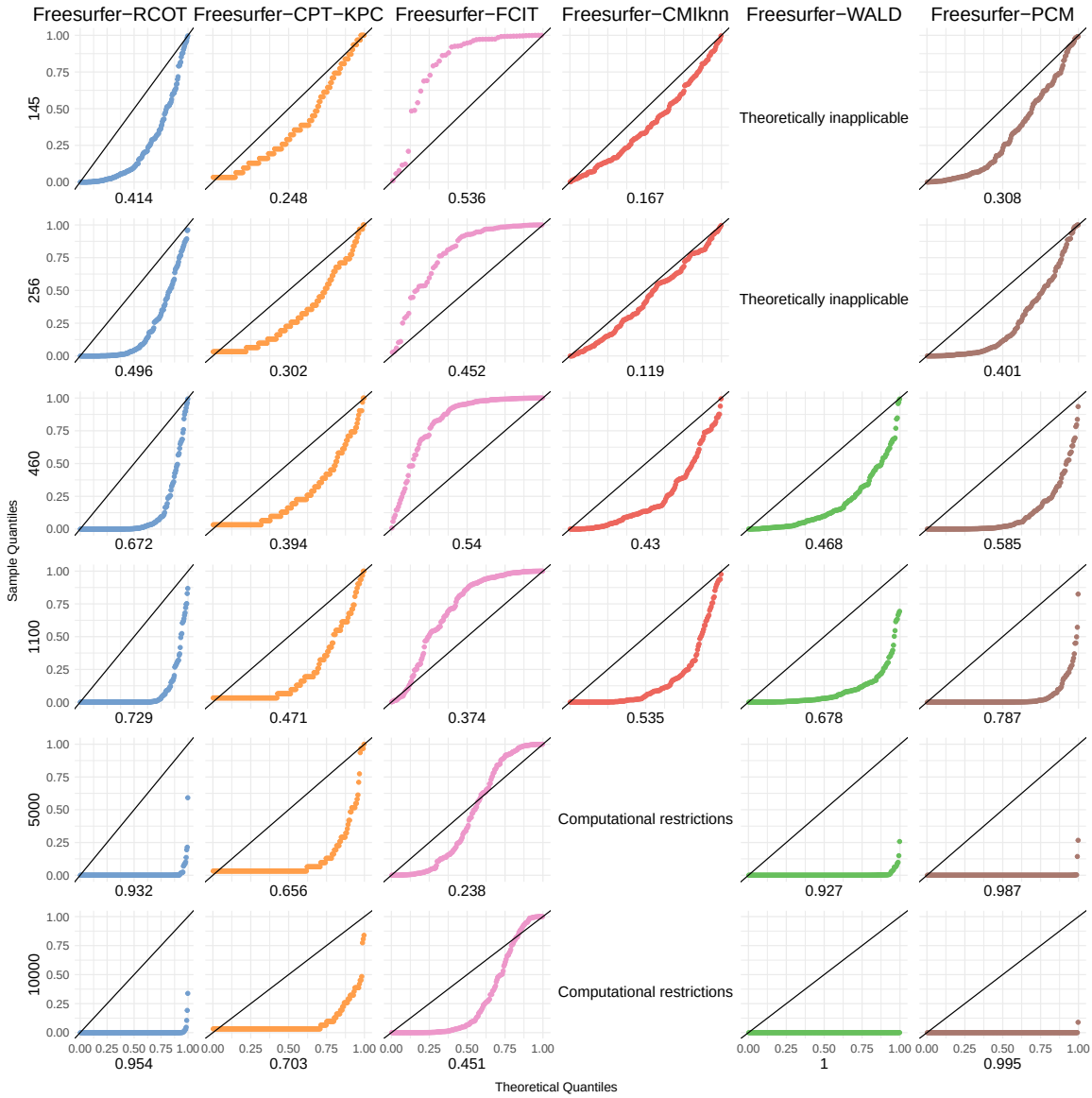

Figure 10: QQ-plots of the observed p-values against the theoretical quantiles of a uniform distribution on the interval $[0, 1]$ for all nonparametric CITs applied to the Freesurfer embedding map at selected sample sizes (indicated at the y-axis label). The results correspond to the DGM under conditional dependence with one confounder and the quadratic confounder relationship $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j \in \mathcal{J}_c})$, where $\mathcal{J}_c$ denotes the index set of continuous variables. The diagonal black lines $y = x$ serve as references for the theoretical quantiles of the uniform distribution on $[0, 1]$. If the p-values are uniformly distributed, they should align along this line. The x-axis labels display the Kolmogorov-Smirnov test statistics used to assess whether the p-values are uniformly distributed; values close to zero indicate a bad calibration of the CIT, as the DGM assumes conditional dependence.

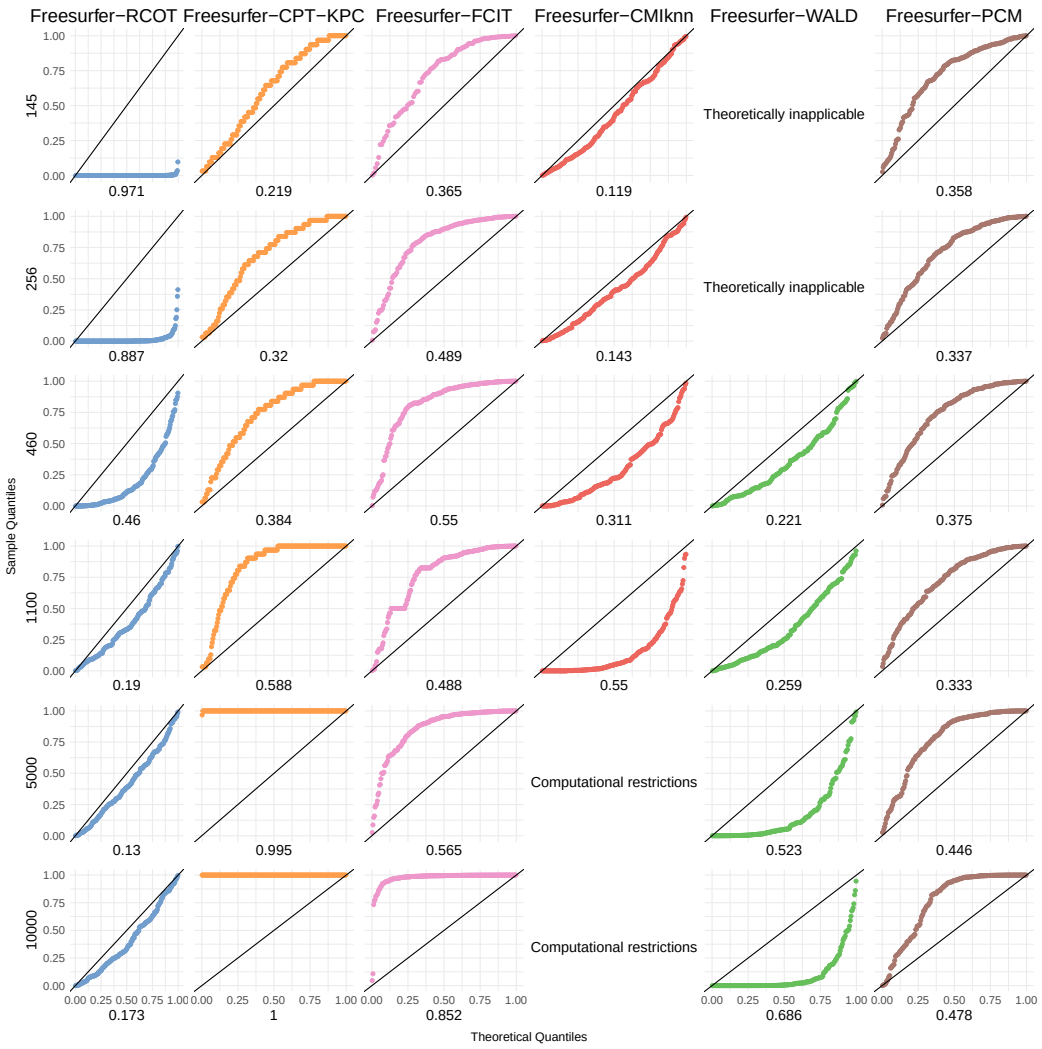

Figure 11: QQ-plots of the observed p-values against the theoretical quantiles of a uniform distribution on the interval $[0, 1]$ for all nonparametric CITs applied to the Freesurfer embedding map at selected sample sizes (indicated at the y-axis label). The results correspond to the DGM under conditional independence with six confounders and the quadratic confounder relationship $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j \in \mathcal{J}_c})$, where $\mathcal{J}_c$ denotes the index set of continuous variables. The diagonal black lines $y = x$ serve as references for the theoretical quantiles of the uniform distribution on $[0, 1]$. If the p-values are uniformly distributed, they should align along this line. The x-axis labels display the Kolmogorov-Smirnov test statistics used to assess whether the p-values are uniformly distributed; smaller values indicate a better calibration of the CIT, as the DGM assumes conditional independence. The `Freesurfer-FCIT` can produce errors for sample sizes up to 460 due to the implementation of its prediction models, resulting in fewer p-values in the corresponding grids.

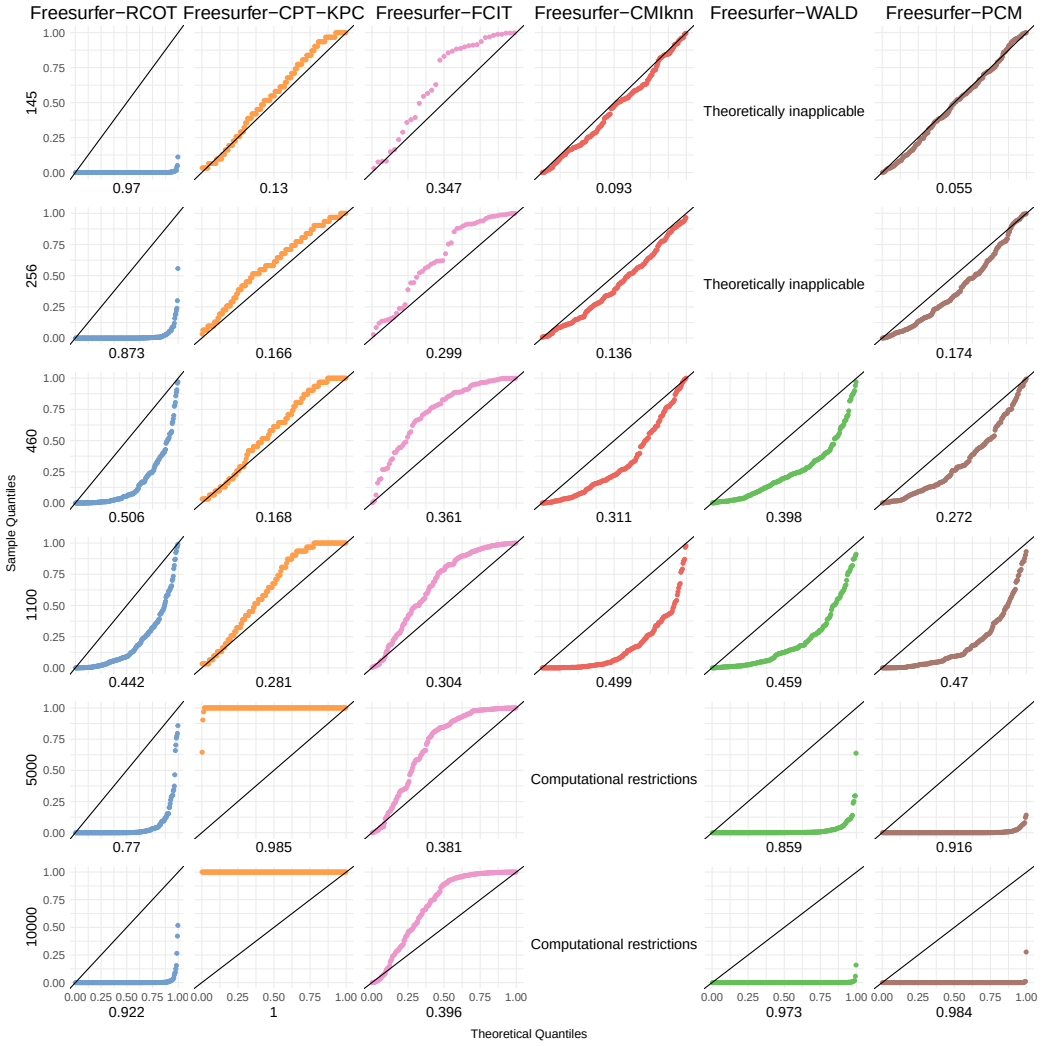

Figure 12: QQ-plots of the observed p-values against the theoretical quantiles of a uniform distribution on the interval $[0, 1]$ for all nonparametric CITs applied to the Freesurfer embedding map at selected sample sizes (indicated at the y-axis label). The results correspond to the DGM under conditional dependence with six confounders and the quadratic confounder relationship $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j \in \mathcal{J}_c})$, where $\mathcal{J}_c$ denotes the index set of continuous variables. The diagonal black lines $y = x$ serve as references for the theoretical quantiles of the uniform distribution on $[0, 1]$. If the p-values are uniformly distributed, they should align along this line. The x-axis labels display the Kolmogorov-Smirnov test statistics used to assess whether the p-values are uniformly distributed; values close to zero indicate a bad calibration of the CIT, as the DGM assumes conditional dependence. The `Freesurfer-FCIT` can produce errors for sample sizes up to 460 due to the implementation of its prediction models, resulting in fewer p-values in the corresponding grids.

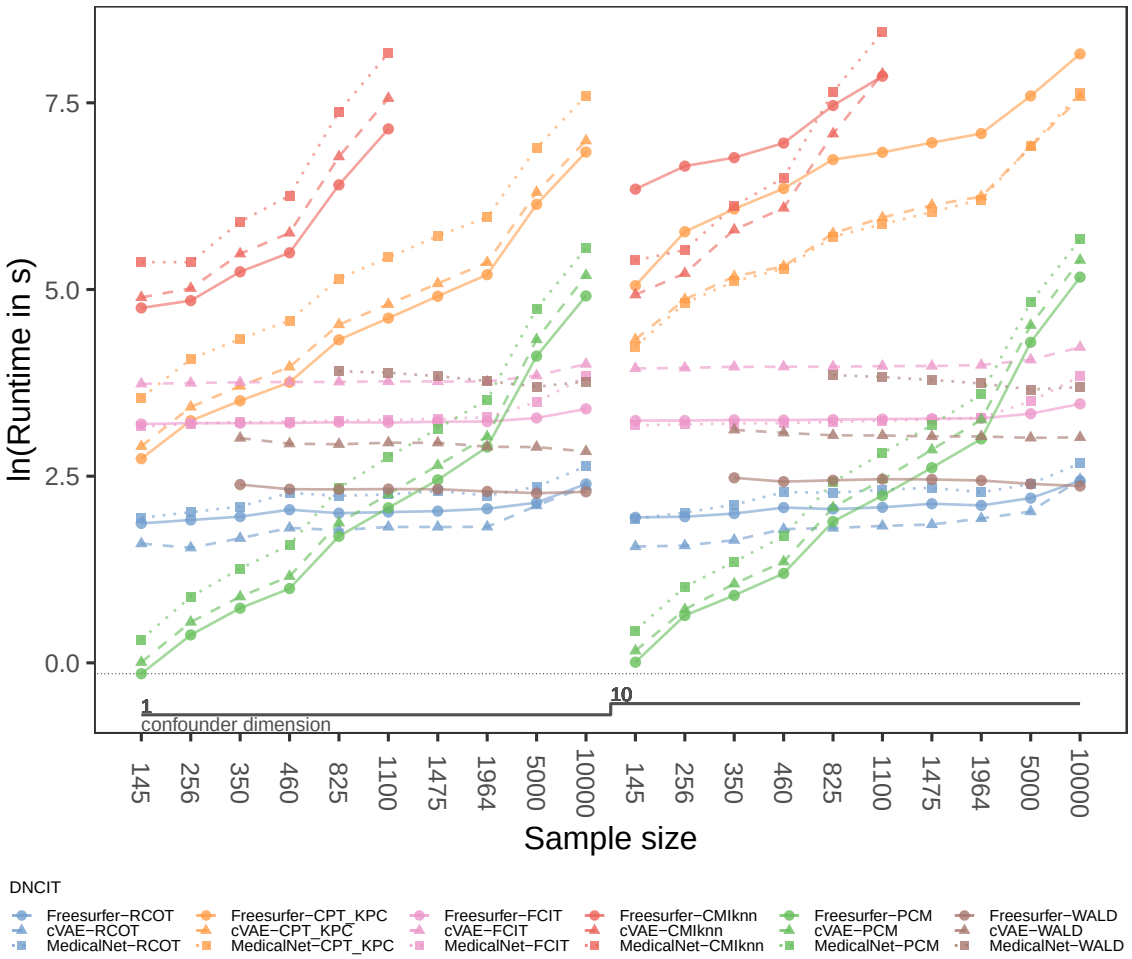

Figure 13: The average of the logarithm of the runtime over 200 seeds of the nonparametric CITs for $c = 0$ (CI) for increasing confounder dimension (columns). Each column increases the sample size from left to right. The confounder relationship is set to $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j \in \mathcal{J}_c})$, where $\mathcal{J}_c$ denotes the index set of continuous variables. The runtime to obtain feature representations from available embedding maps is ignored, as the embedding maps were already trained and the forward pass for all observations has to be computed only once before all simulations.

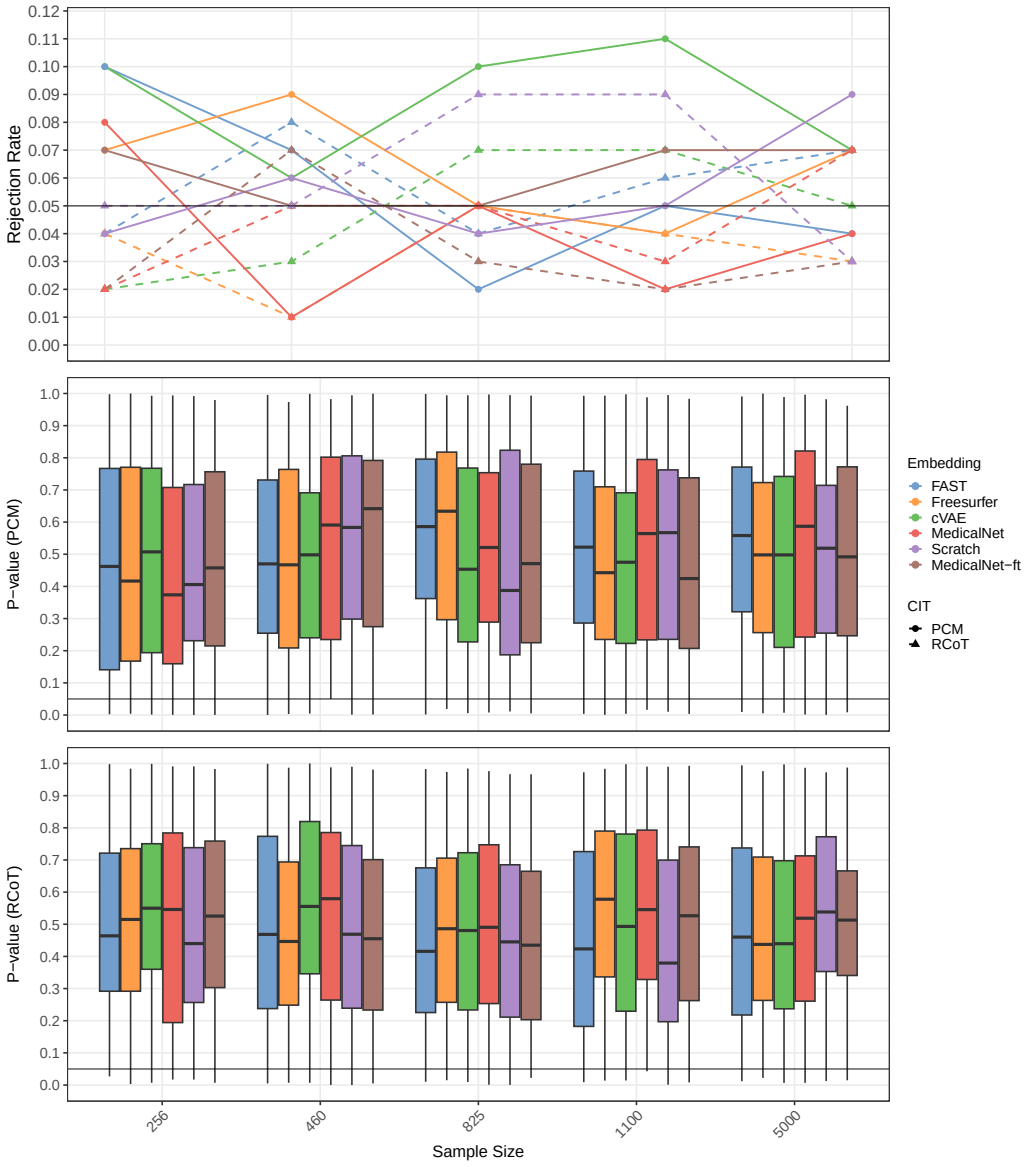

Figure 14: Rejection rates and p-value boxplots over 100 seeds for DNCITs combining the embedding maps and nonparametric CITs specified in the legend. The sample sizes are equidistantly depicted, the p-values using a $\log_{10}$ scaling. The black, solid horizontal lines in the p-value panels mark the level $\alpha = 0.05$ at which a test is rejected. The results correspond to the DGM under conditional independence with $\varepsilon_i \sim N(0, 0.5^2)$, one confounder and the quadratic confounder relationship $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j \in \mathcal{J}_c})$, where $\mathcal{J}_c$ denotes the index set of continuous variables. Rejection rates below and around 0.05 and uniformly distributed p-values on $[0, 1]$ indicate a better performance of the DNCIT, as the DGM assumes conditional independence.

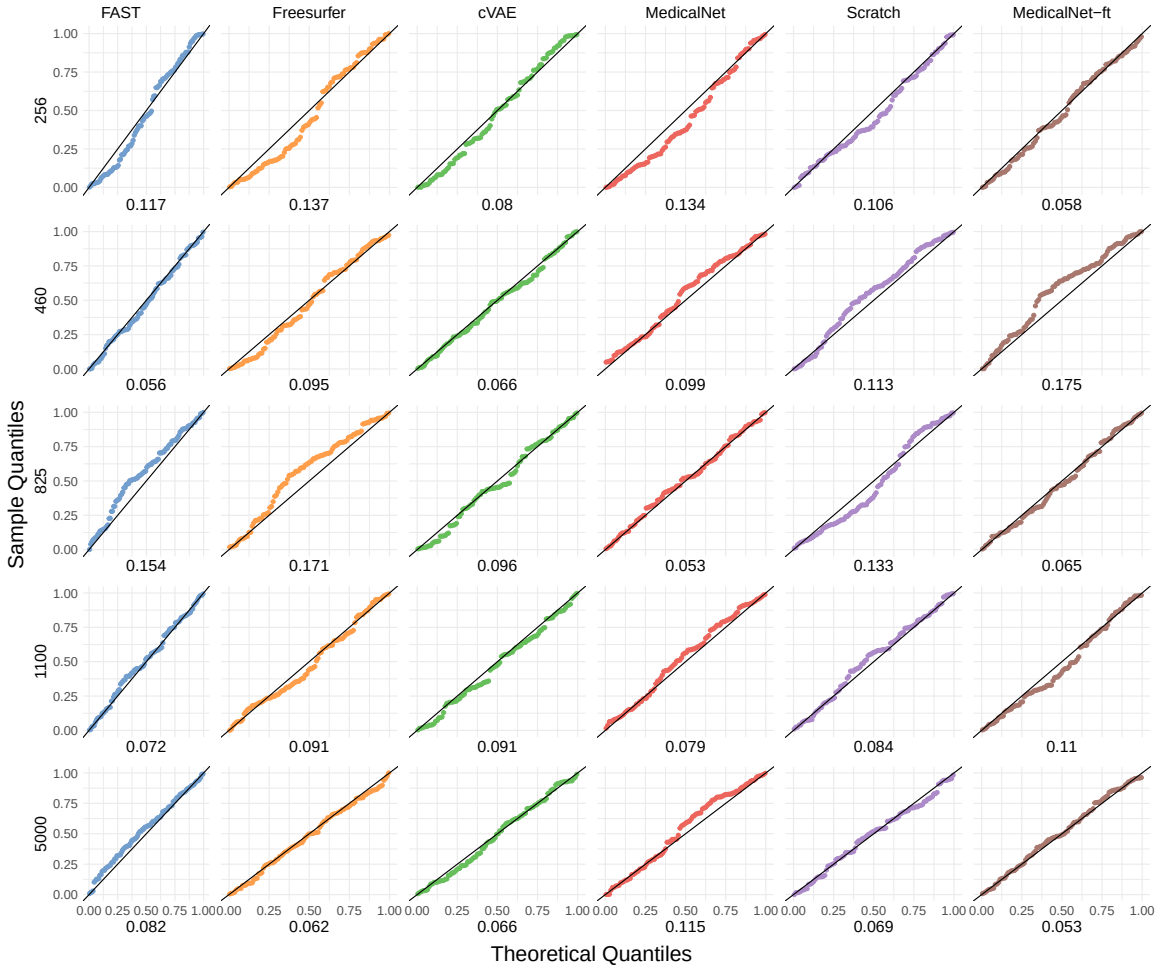

Figure 15: QQ-plots of the observed p-values over 100 seeds against the theoretical quantiles of a uniform distribution on the interval $[0, 1]$ for the PCM applied to the embedding maps FAST, Freesurfer, cVAE, MedicalNet, fine-tuned MedicalNet, and one trained from scratch, at selected sample sizes (indicated at the y-axis label). The results correspond to the DGM under conditional independence with $\varepsilon_i \sim N(0, 0.5^2)$, one confounder and the quadratic confounder relationship $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j \in \mathcal{J}_c})$, where $\mathcal{J}_c$ denotes the index set of continuous variables. The diagonal black lines $y = x$ serve as references for the theoretical quantiles of the uniform distribution on $[0, 1]$. If the p-values are uniformly distributed, they should align along this line. The x-axis labels display the Kolmogorov-Smirnov test statistics used to assess whether the p-values are uniformly distributed; smaller values indicate a better calibration of the CIT, as the DGM assumes conditional independence.

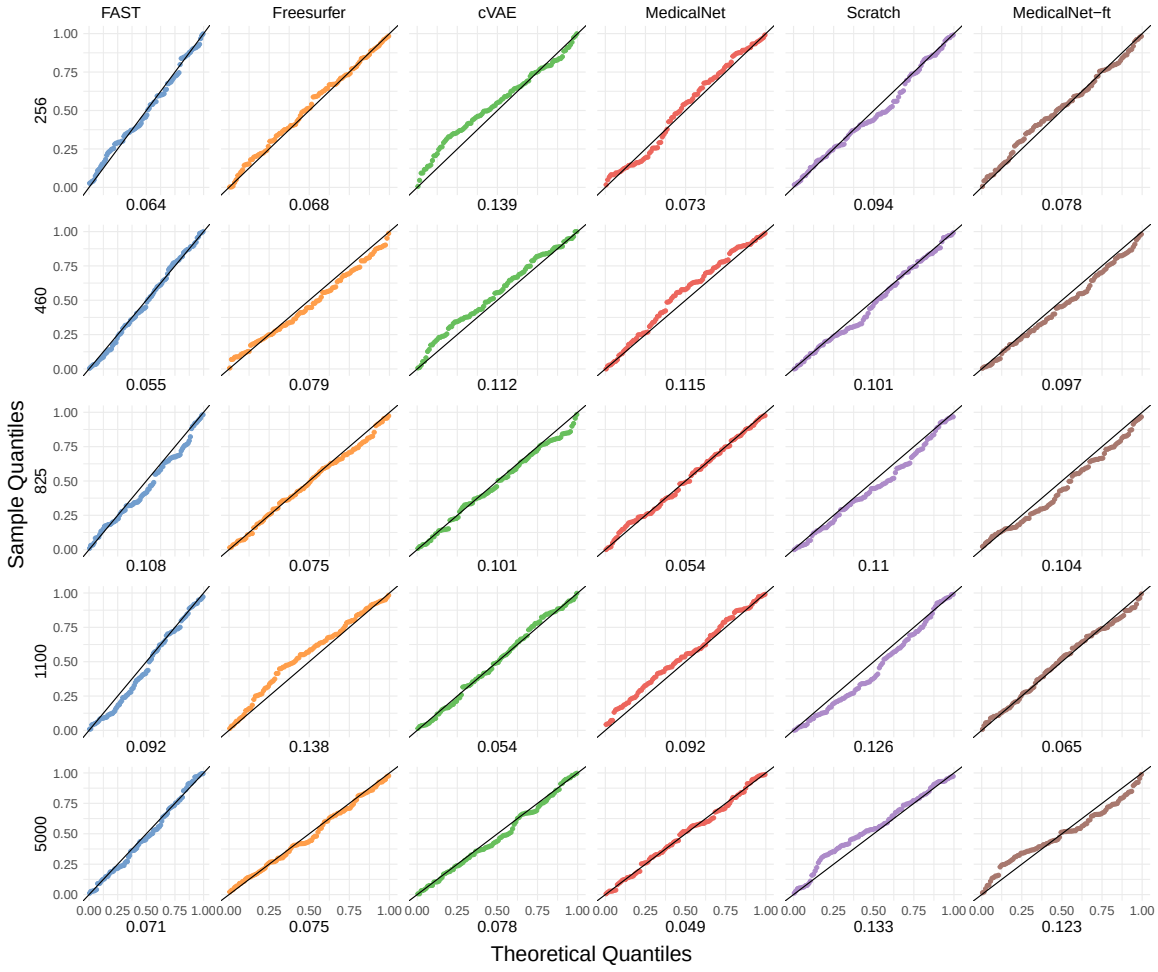

Figure 16: QQ-plots of the observed p-values over 100 seeds against the theoretical quantiles of a uniform distribution on the interval $[0, 1]$ for the RCoT applied to the embedding maps FAST, Freesurfer, cVAE, MedicalNet, fine-tuned MedicalNet, and one trained from scratch, at selected sample sizes (indicated at the y-axis label). The results correspond to the DGM under conditional independence with $\varepsilon_i \sim N(0, 0.5^2)$, one confounder and the quadratic confounder relationship $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j \in \mathcal{J}_c})$, where $\mathcal{J}_c$ denotes the index set of continuous variables. The diagonal black lines $y = x$ serve as references for the theoretical quantiles of the uniform distribution on $[0, 1]$. If the p-values are uniformly distributed, they should align along this line. The x-axis labels display the Kolmogorov-Smirnov test statistics used to assess whether the p-values are uniformly distributed; smaller values indicate a better calibration of the CIT, as the DGM assumes conditional independence.

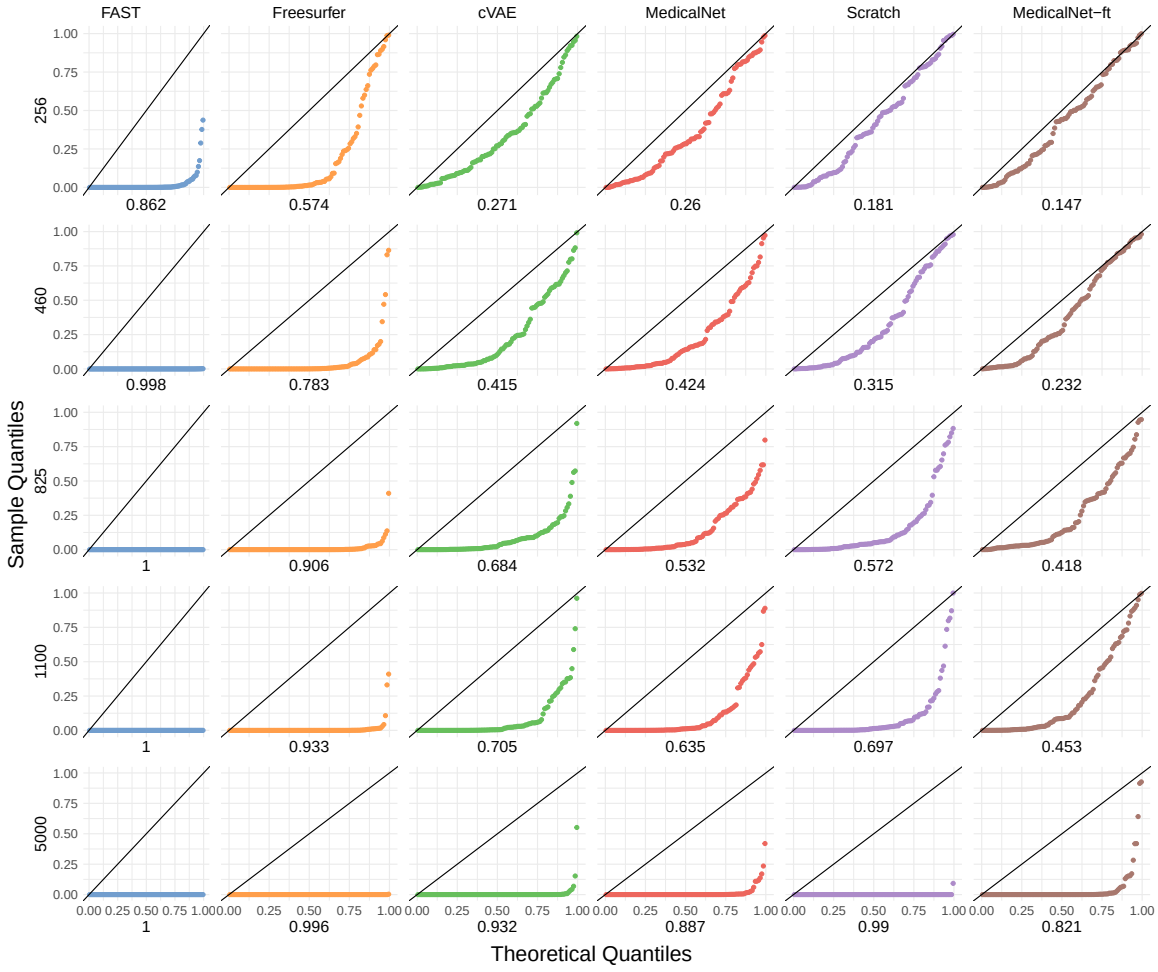

Figure 17: QQ-plots of the observed p-values over 100 seeds against the theoretical quantiles of a uniform distribution on the interval $[0, 1]$ for the PCM applied to the embedding maps FAST, Freesurfer, cVAE, MedicalNet, fine-tuned MedicalNet, and one trained from scratch, at selected sample sizes (indicated at the y-axis label). The results correspond to the DGM under conditional dependence with $\varepsilon_i \sim N(0, 0.5^2)$, one confounder and the quadratic confounder relationship $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j \in \mathcal{J}_c})$, where $\mathcal{J}_c$ denotes the index set of continuous variables. The diagonal black lines $y = x$ serve as references for the theoretical quantiles of the uniform distribution on $[0, 1]$. If the p-values are uniformly distributed, they should align along this line. The x-axis labels display the Kolmogorov-Smirnov test statistics used to assess whether the p-values are uniformly distributed; values close to zero indicate a bad calibration of the CIT, as the DGM assumes conditional dependence.

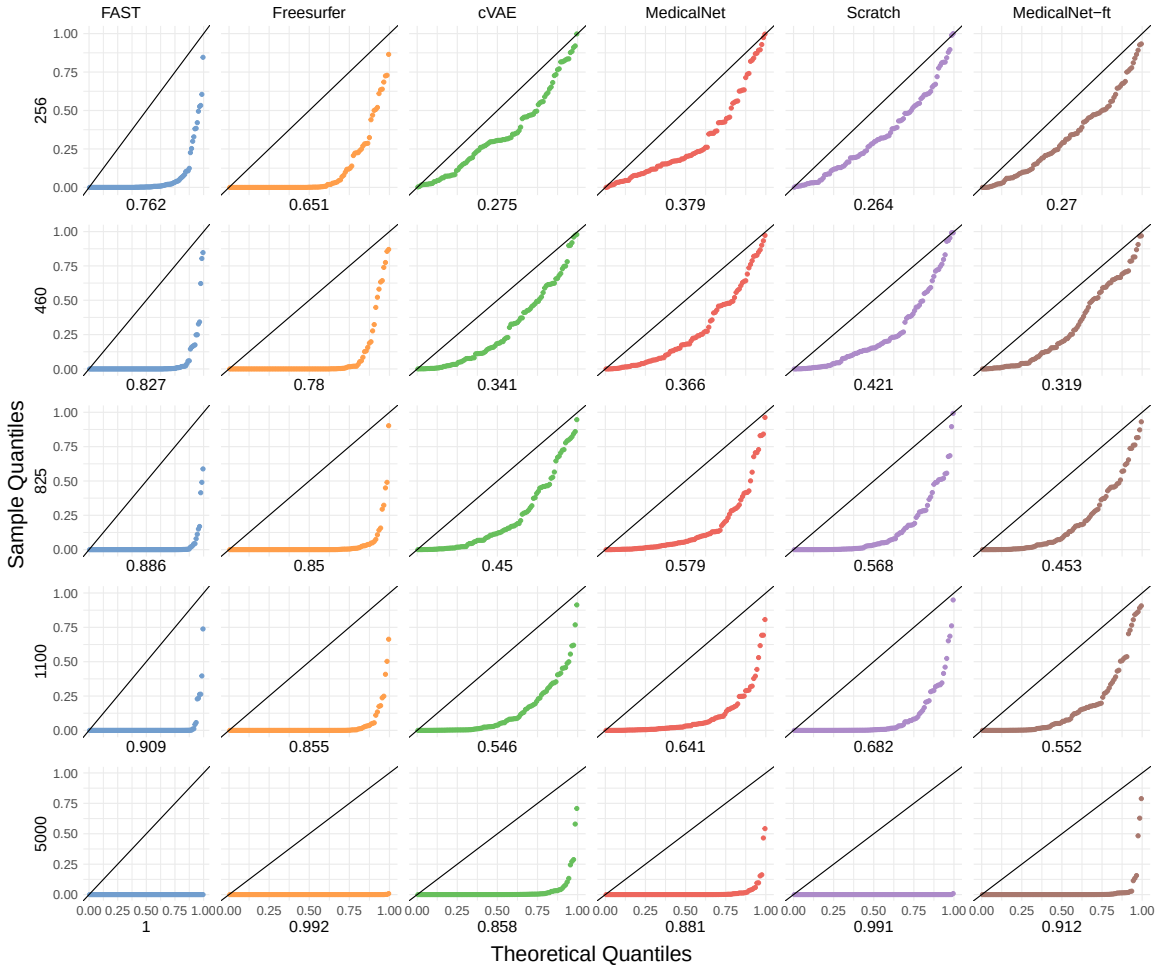

Figure 18: QQ-plots of the observed p-values over 100 seeds against the theoretical quantiles of a uniform distribution on the interval $[0, 1]$ for the RCoT applied to the embedding maps FAST, Freesurfer, cVAE, MedicalNet, fine-tuned MedicalNet, and one trained from scratch, at selected sample sizes (indicated at the y-axis label). The results correspond to the DGM under conditional dependence with $\varepsilon_i \sim N(0, 0.5^2)$, one confounder and the quadratic confounder relationship $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j \in \mathcal{J}_c})$, where $\mathcal{J}_c$ denotes the index set of continuous variables. The diagonal black lines $y = x$ serve as references for the theoretical quantiles of the uniform distribution on $[0, 1]$. If the p-values are uniformly distributed, they should align along this line. The x-axis labels display the Kolmogorov-Smirnov test statistics used to assess whether the p-values are uniformly distributed; values close to zero indicate a bad calibration of the CIT, as the DGM assumes conditional dependence.

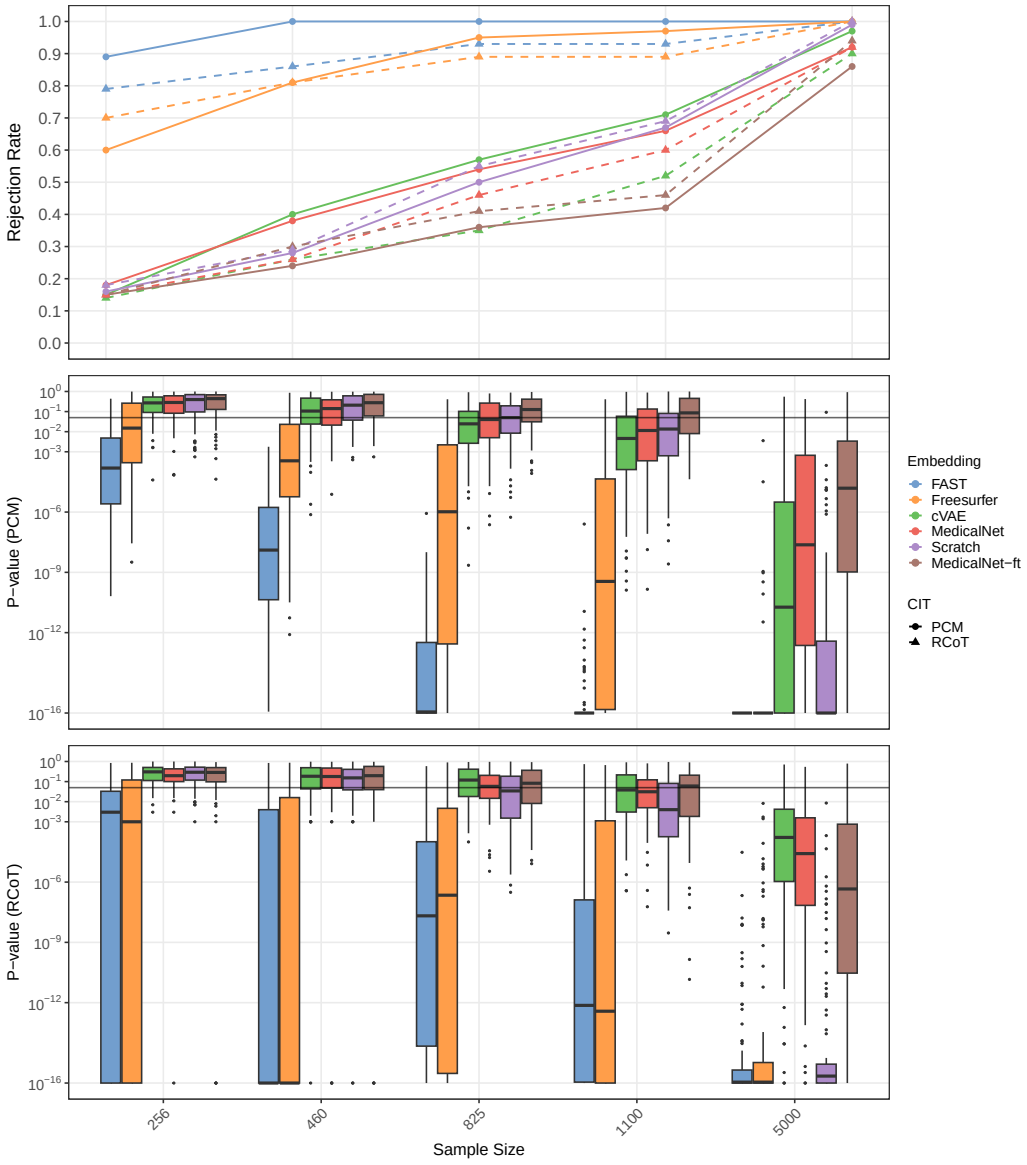

Figure 19: Rejection rates and p-value boxplots over 100 seeds for DNCITs combining the embedding maps and nonparametric CITs specified in the legend. The sample sizes are equidistantly depicted, the p-values using a $\log_{10}$ scaling. P-values lower than 10^{-16} are set to 10^{-16}. The black, solid horizontal lines in the p-value panels mark the level $\alpha = 0.05$ at which a test is rejected. The results correspond to the DGM under conditional dependence with $\varepsilon_i \sim N(0, 0.5^2)$, one confounder and the quadratic confounder relationship $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j \in \mathcal{J}_c})$, where $\mathcal{J}_c$ denotes the index set of continuous variables. Higher rejection rates and lower p-values indicate a better performance of the DNCIT, as the DGM assumes conditional dependence.

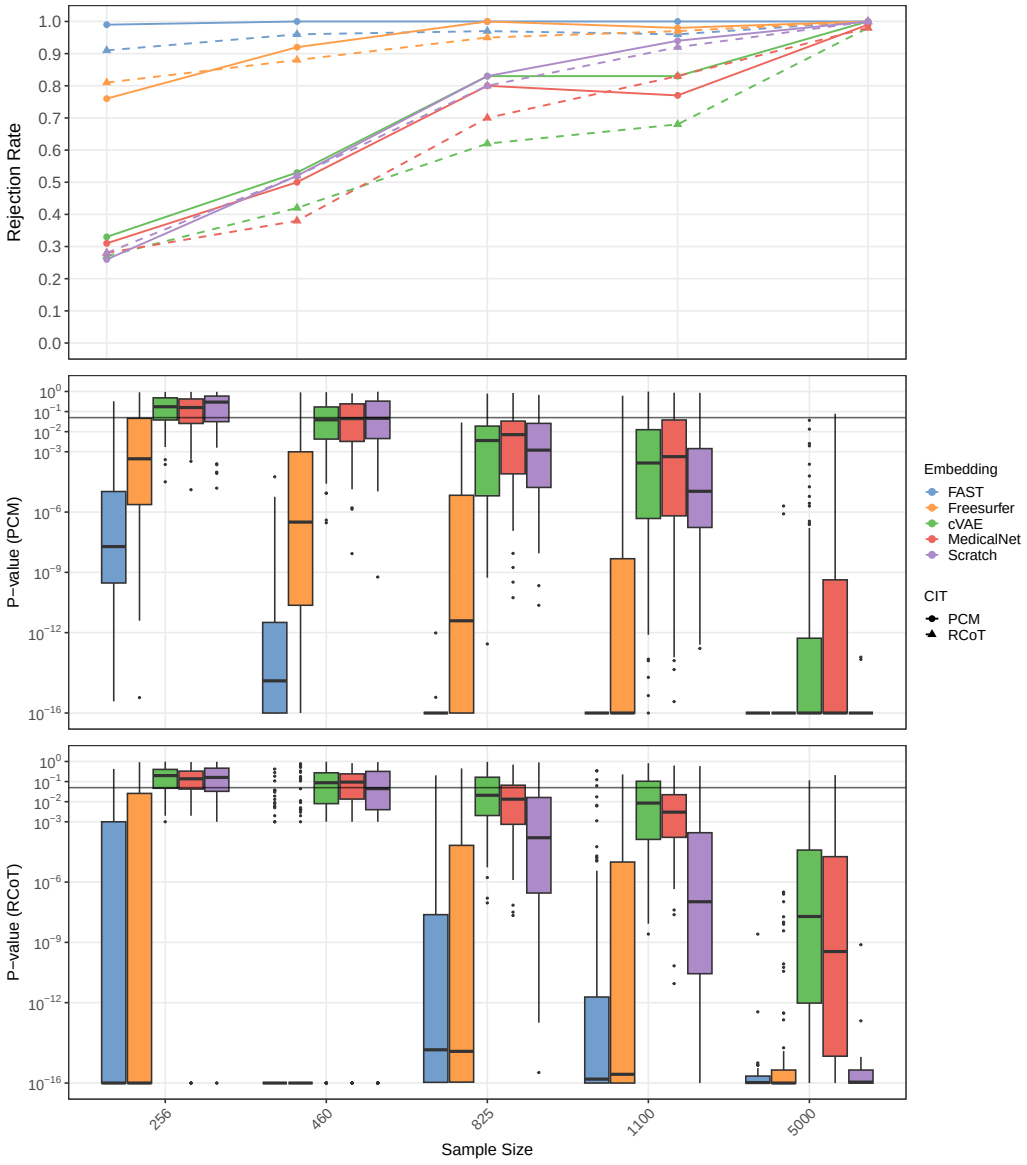

Figure 20: Rejection rates and p-value boxplots over 100 seeds for DNCITs combining the embedding maps and nonparametric CITs specified in the legend. The sample sizes are equidistantly depicted, the p-values using a $\log_{10}$ scaling. P-values lower than 10^{-16} are set to 10^{-16}. The black, solid horizontal lines in the p-value panels mark the level $\alpha = 0.05$ at which a test is rejected. The results correspond to the DGM under conditional dependence with $\varepsilon_i \sim N(0, 0.1^2)$, one confounder and the quadratic confounder relationship $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j \in \mathcal{J}_c})$, where $\mathcal{J}_c$ denotes the index set of continuous variables. Higher rejection rates and lower p-values indicate a better performance of the DNCIT, as the DGM assumes conditional dependence.

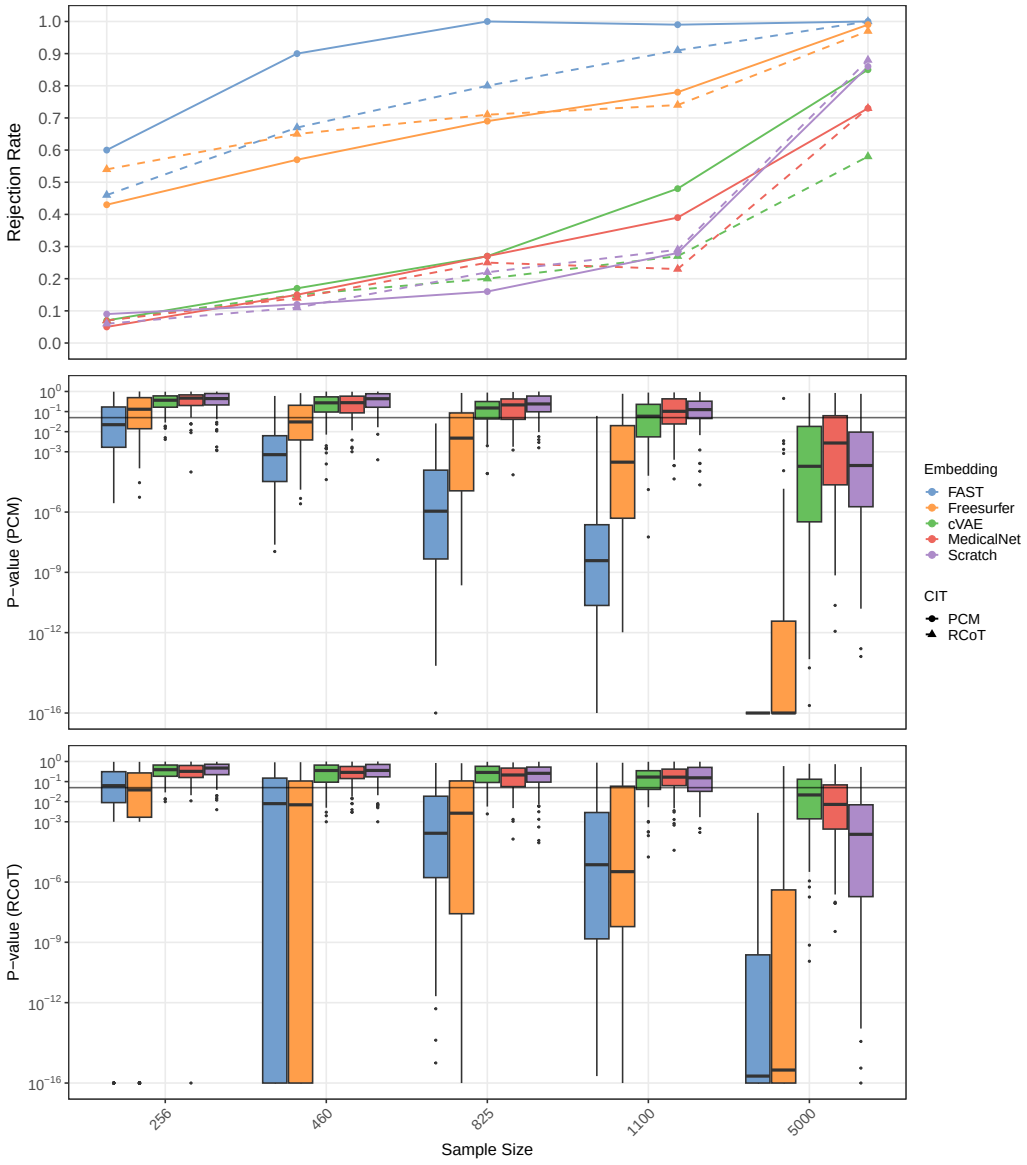

Figure 21: Rejection rates and p-value boxplots over 100 seeds for DNCITs combining the embedding maps and nonparametric CITs specified in the legend. The sample sizes are equidistantly depicted, the p-values using a $\log_{10}$ scaling. P-values lower than 10^{-16} are set to 10^{-16}. The black, solid horizontal lines in the p-value panels mark the level $\alpha = 0.05$ at which a test is rejected. The results correspond to the DGM under conditional dependence with $\varepsilon_i \sim N(0,1)$, one confounder and the quadratic confounder relationship $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j\in\mathcal{J}_c})$, where $\mathcal{J}_c$ denotes the index set of continuous variables. Higher rejection rates and lower p-values indicate a better performance of the DNCIT, as the DGM assumes conditional dependence.

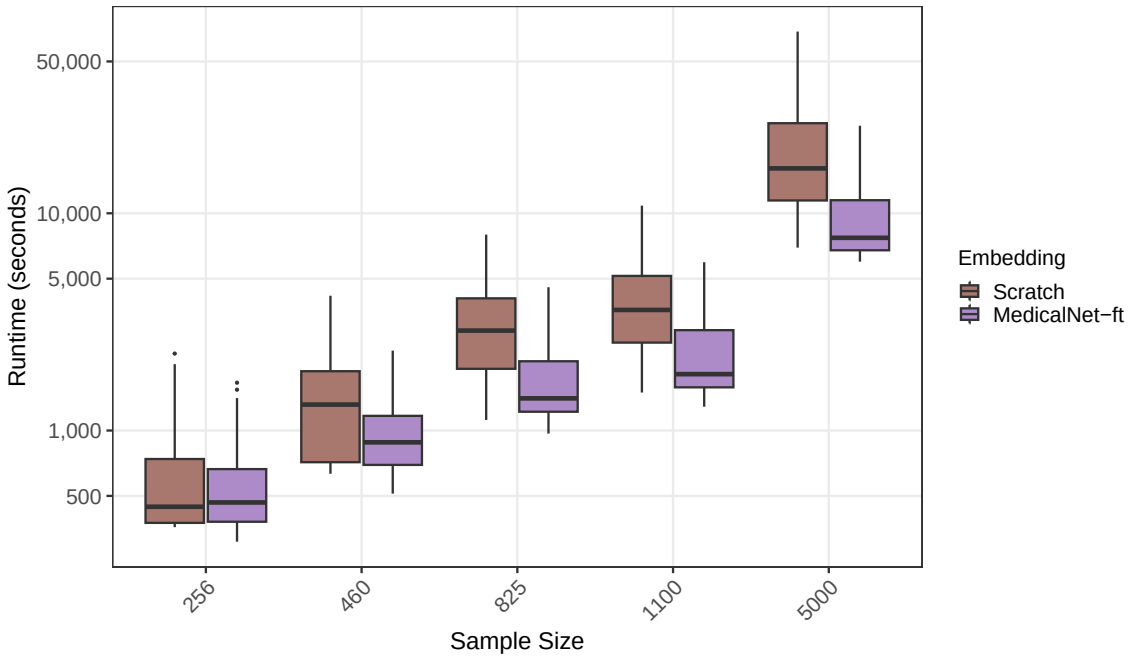

Figure 22: The runtime using a $\log_{10}$ scaling of the from scratch trained and the fine-tuned MedicalNet embedding map for $c = 0$ (CI) and $\varepsilon_i \sim N(0, 0.5^2)$ over 100 seeds. Each column increases the sample size from left to right. The confounder is set to age and the confounder relationship is set to $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j \in \mathcal{J}_c})$, where $\mathcal{J}_c$ denotes the index set of continuous variables.

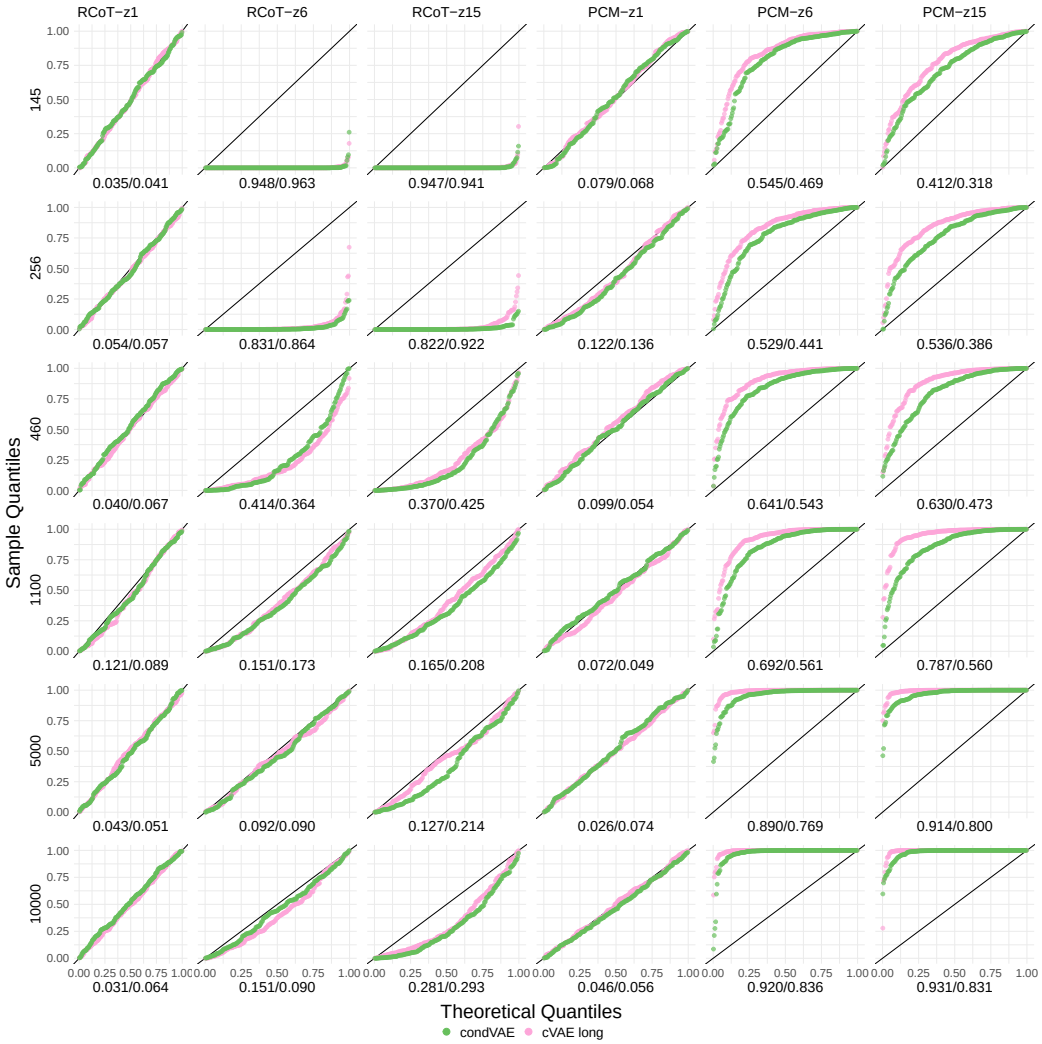

Figure 23: QQ-plots of the observed p-values over 200 seeds against the theoretical quantiles of a uniform distribution on the interval $[0, 1]$ for the RCoT and the PCM applied to the embedding maps cVAE and the penultimate layer of the cVAE at selected sample sizes (indicated at the y-axis label). The results correspond to the DGM under conditional independence with $\varepsilon_i \sim N(0, 1)$, one, six and fifteen confounder (x-axis) and the quadratic confounder relationship $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j \in \mathcal{J}_c})$, where $\mathcal{J}_c$ denotes the index set of continuous variables. The diagonal black lines $y = x$ serve as references for the theoretical quantiles of the uniform distribution on $[0, 1]$. If the p-values are uniformly distributed, they should align along this line. The x-axis labels display the Kolmogorov-Smirnov test statistics (left for the penultimate layer of the cVAE/ right for the cVAE) used to assess whether the p-values are uniformly distributed; smaller values indicate a better calibration of the CIT, as the DGM assumes conditional independence.

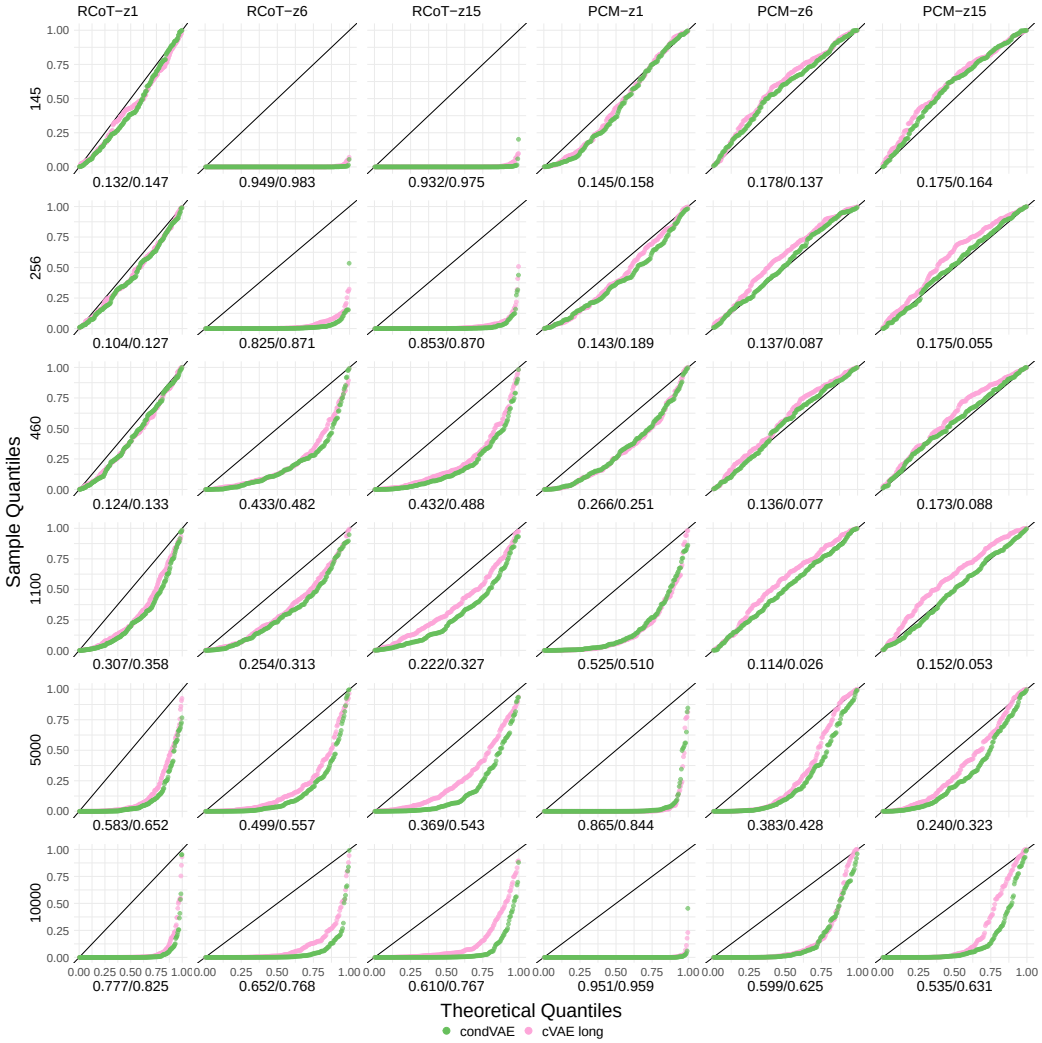

Figure 24: QQ-plots of the observed p-values over 200 seeds against the theoretical quantiles of a uniform distribution on the interval $[0, 1]$ for the RCoT and the PCM applied to the embedding maps cVAE and the penultimate layer of the cVAE at selected sample sizes (indicated at the y-axis label). The results correspond to the DGM under conditional dependence with $\varepsilon_i \sim N(0, 1)$, one, six and fifteen confounder (x-axis) and the quadratic confounder relationship $g_z(\mathbf{s}) = (\mathbf{s}^\top, (s_j^2)_{j \in \mathcal{J}_c})$, where $\mathcal{J}_c$ denotes the index set of continuous variables. The diagonal black lines $y = x$ serve as references for the theoretical quantiles of the uniform distribution on $[0, 1]$. If the p-values are uniformly distributed, they should align along this line. The x-axis labels display the Kolmogorov-Smirnov test statistics (left for the penultimate layer of the cVAE/ right for the cVAE) used to assess whether the p-values are uniformly distributed; values close to zero indicate a bad calibration of the CIT, as the DGM assumes conditional dependence.

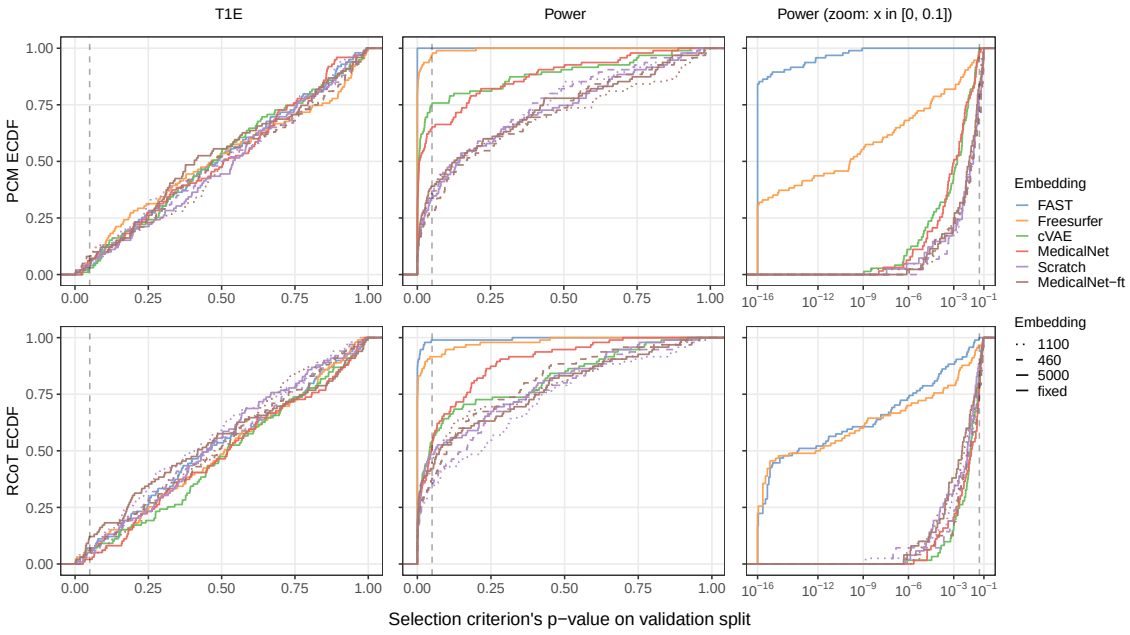

Figure 25: The ECDFs of the p-values of the embedding selection criterion using the PCM (top row) and RCoT (bottom row) applied to different embedding maps under the null hypothesis (left) and alternative hypothesis (center, right). The line types indicate the sample sizes used for the DNCITs for the trained embedding maps (Scratch and MedicalNet-ft). The fixed, non-DNCIT-specific embedding maps are fixed across sample sizes. P-values smaller than 10^{-16} are set to 10^{-16}.

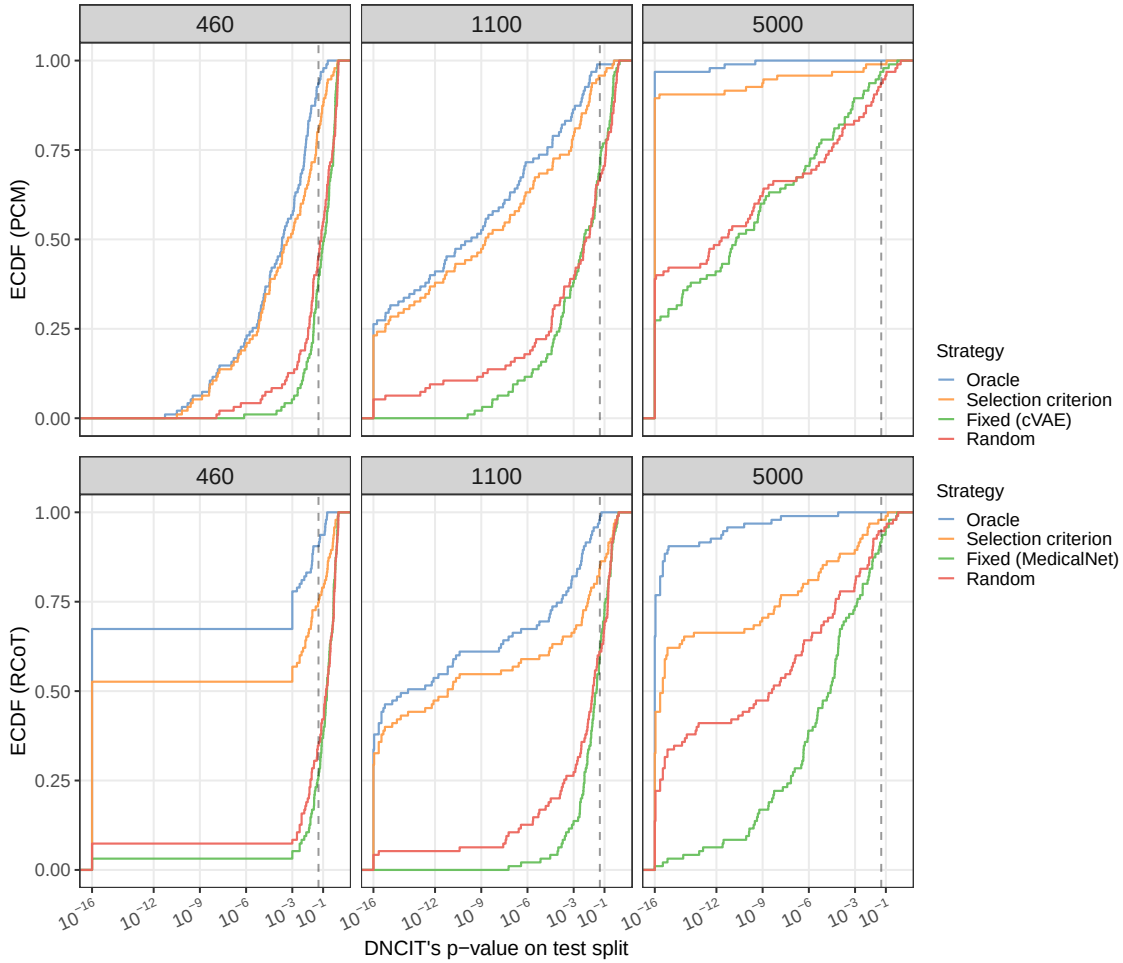

Figure 26: Empirical cumulative distribution functions (ECDFs) of the p-values of the DNCITs using the PCM (top row) and RCoT (bottom row) and applying a given selection strategy for the embedding map (indicated by color). For a given seed, the *oracle* selects the embedding map with the lowest p-value on the test set among Freesurfer, cVAE, MedicalNet, Scratch and MedicalNet-ft embedding maps, the *criterion* selects the one with the smallest criterion value, i.e. p-value on the validation split, *fixed* selects the cVAE (top row) and MedicalNet (bottom row), and *random* draws randomly from the embedding maps. Sample sizes of test sets used for the DNCITs are given in the panel titles. For DNCIT-specific embedding maps, only half of the validation sample is used to obtain the p-value for the selection criterion (cf. also Subsection 3.2).

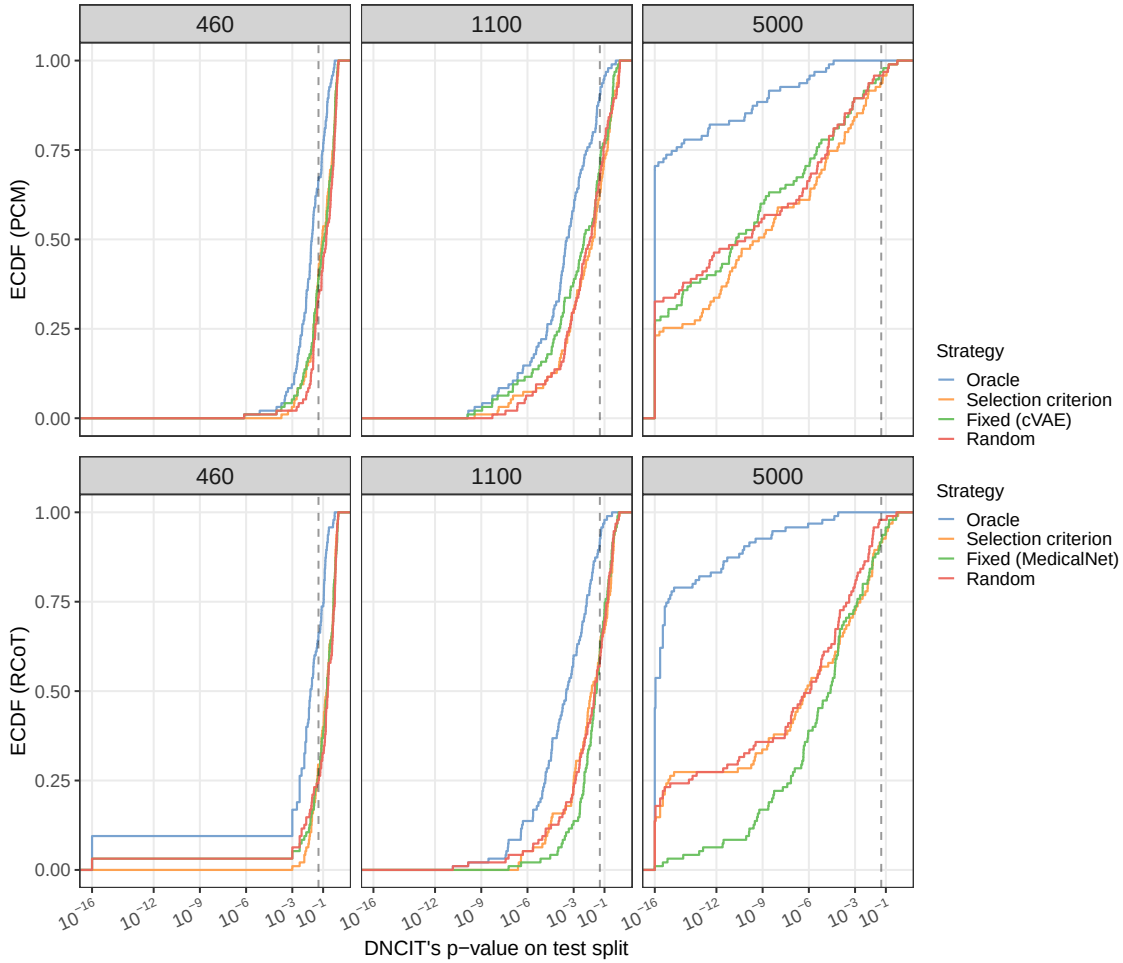

Figure 27: Empirical cumulative distribution functions (ECDFs) of the p-values of the DNCITs using the PCM (top row) and RCoT (bottom row) and applying a given selection strategy for the embedding map (indicated by color). For a given seed, the *oracle* selects the embedding map with the lowest p-value on the test set among cVAE, MedicalNet, Scratch and MedicalNet-ft embedding maps, the *criterion* selects the one with the smallest criterion value, i.e. p-value on the validation split, *fixed* selects the cVAE (top row) and MedicalNet (bottom row), and *random* draws randomly from the embedding maps. Sample sizes of test sets used for the DNCITs are given in the panel titles. For DNCIT-specific embedding maps, only half of the validation sample is used to obtain the p-value for the selection criterion (cf. also Subsection 3.2).

Appendix D. Details, Adaptation and Implementation of Selected Conditional Independence Tests

First, we give a short summary of the selected nonparametric CITs and analyze them afterwards in more detail w.r.t. our setting. We construct a nonparametric CIT using the KPC from Huang et al. (2022) as the test statistic together with the CPT from Berrett et al. (2020); Spisak (2022). While the kernel of the KPC adapts well to the dimension of the feature representations and its geometric graph functionals to the dimension of the confounders, the CPT framework ensures the validity of the test. In addition, the KPC allows a flexible choice of kernel and is implemented in a stable manner. Finally, ancillary data for $\mathbb{P}^{Y|Z}$ is often available, especially in the context of medical imaging, which can be expensive for many patients, while tabular measurements may be more readily available in larger quantities, such as in the UKB (Sudlow et al., 2015). In addition, we select the RCoT of Strobl et al. (2019) because of its fast approximation of the KCIT with similar performance, its stable implementation, and the often strong performance of kernel-based methods in high-dimensional settings. Furthermore, we choose the prediction-based CITs FCIT of Chalupka et al. (2018) because of its adaptation to the different dimension between the feature representations of the image X and the scalar Y. Moreover, we select the PCM (Lundborg et al., 2024). Like the FCIT, it is also based on several tabular regression models. However, it includes a debiasing step for the regressions and provides stronger theoretical results as well as a more flexible implementation (Kook and Lundborg, 2024). Finally, the CMIknn of Runge (2018) is chosen to represent metric-based CITs.

D.1 The Conditional Permutation Test and the Kernel Partial Coefficient

Throughout our empirical results, we approximate for CPT-based CITs $\mathbb{P}^{Y|Z_i=\mathbf{z}_i}$ by a normal distribution with mean and variance estimated by a generalized additive model in sample, where the effect of each continuous confounder in $Z_i = (Z_{i1}, \dots, Z_{ip})$ is modeled by a smooth term and that of each categorical confounder with a separate parameter for each category.[2] This is analogous to the approach in Spisak (2022), however we use in the implementation the more reliable R package mgcv from Wood (2015) compared to the Python package pyGAM from Servén and Brummitt (2018) used in Spisak (2022). Nevertheless, the method to approximate $\mathbb{P}^{Y|Z=\mathbf{z}_i}$ is adaptable and depends on $\mathcal{Z}$ and $\mathcal{Y}$ as well as on potential supplementary data $(Y_i, Z_i)_{i=n+1,\dots,n+n'}$.

The time complexity of CPT-based CITs is the time complexity of approximating $\mathbb{P}^{Y|Z}$ plus M times the time complexity of estimating the test statistic, where M is the number of permutations of Y. For example, a detailed description of the time complexity for generalized additive models and their extensions can be found in Wood (2020), and in our specific implementation is given by $\mathcal{O}(ns^2)$, where s is the number of parameters in the basis representation of Z used to approximate $\mathbb{P}^{Y|Z}$ and n is the sample size of the data set. However, this could be reduced using different implementations, and we observe throughout the simulation study that for the KPC test statistic that we consider, estimating the test statistic is the bottleneck in terms of computational cost.

2. Other distribution families are available for generalized additive models and could be used instead of the normal distribution. Additionally, modeling continuous and categorical confounders could be made more flexible.

Together with the CPT, we have chosen the conditional dependence measure recently presented in Huang et al. (2022) as test statistic, which extends upon measures established in Deb et al. (2020); Azadkia and Chatterjee (2021). Specifically, they propose the kernel partial correlation (KPC) as

$$KPC_\theta((X^\omega, Y, Z)) \coloneqq \frac{\mathbb{E}[MMD^2(\mathbb{P}^{X^\omega|Y,Z}, \mathbb{P}^{X^\omega|Z})]}{\mathbb{E}[MMD^2(\delta_{X^\omega}, \mathbb{P}^{X^\omega|Z})]} \tag{9}$$

where MMD denotes the maximum mean discrepancy as for example defined in Gretton et al. (2012), δ_{X^ω} is the Dirac measure at X^ω and θ are the parameters of the kernels introduced below.

For a characteristic, finite kernel $k_{\mathfrak{X}}$, separable RKHS $\mathcal{H}_{\mathfrak{X}}$, and the assumption that X^ω is not a measurable function of Z, it holds that $KPC_\theta((X^\omega, Y, Z)) \in [0, 1]$ and in particular, $KPC_\theta((X^\omega, Y, Z)) = 0$ if and only if $X^\omega \perp\!\!\!\perp Y|Z$. Thus, together with the CPT, the hypothesis (2) can be tested, under the above assumptions equivalently, by testing

$$H^\omega_{0,KPC} : KPC_\theta((X^\omega, Y, Z)) = 0 \quad vs. \quad H^\omega_{1,KPC} : KPC_\theta((X^\omega, Y, Z)) > 0.$$

Furthermore, the authors showed in Huang et al. (2022, Lemma 1,2) that the KPC is well defined and equivalent to

$$KPC_\theta((X^\omega, Y, Z)) = \frac{\mathbb{E}[\mathbb{E}[k_{\mathfrak{X}}(X_2^\omega, X_2^{\omega'})|Y, Z]] - \mathbb{E}[\mathbb{E}[k_{\mathfrak{X}}(X_1^\omega, X_1^{\omega'})|Z]]}{\mathbb{E}[k_{\mathfrak{X}}(X^\omega, X^\omega)] - \mathbb{E}[\mathbb{E}[k_{\mathfrak{X}}(X_1^\omega, X_1^{\omega'})|Z]]}$$

where the joint distributions of $(Z, X_1^\omega, X_1^{\omega'})$ and $(Z, X_2^\omega, X_2^{\omega'}, Y)$, respectively, are given by

$$\begin{aligned} &Z \sim \mathbb{P}^Z, \quad X_1^\omega|Z \sim \mathbb{P}^{X^\omega|Z}, \quad X_1^{\omega'}|Z \sim \mathbb{P}^{X^\omega|Z}, \quad X_1^\omega \perp\!\!\!\perp X_1^{\omega'}|Z; \\ &(Y, Z) \sim \mathbb{P}^{Y,Z}, \quad X_2^\omega|Y, Z \sim \mathbb{P}^{X^\omega|Y,Z}, \quad X_2^{\omega'}|Y, Z \sim \mathbb{P}^{X^\omega|Y,Z}, \quad X_2^\omega \perp\!\!\!\perp X_2^{\omega'}|X, Z. \end{aligned}$$

Then, the authors propose a graph-based estimator in Huang et al. (2022, sec. 3). Specifically, denote $\ddot{Y} = (Y, Z)$ on $\mathcal{Y} \times \mathcal{Z}$, and let G_n^Z and $G_n^{\ddot{Y}}$ be graphs of a geometric graph functional, such as knn graphs, on $\mathcal{Z}$ and $\mathcal{Y} \times \mathcal{Z}$, respectively, with nodes $Z^n = (Z_1, \ldots, Z_n)$ and $\ddot{Y}^n$, respectively. Moreover, let $\mathcal{E}(G_n^Z)$ and $\mathcal{E}(G_n^{\ddot{Y}})$ denote the corresponding edge sets and d_i^Z and $d_i^{\ddot{Y}}$ be the degrees of Z_i and $\ddot{Y}_i$, $i = 1, \ldots, n$. Then, the estimator of the KPC (9) called graph KPC is defined as in Huang et al. (2022, sec. 3.1) by

$$T^\omega_{\theta_{KPC}}((X^\omega, Y, Z)^n) = \frac{\frac{1}{n}\sum_{i=1}^n \frac{1}{d_i^{\ddot{Y}}} \sum_{j:(i,j)\in\mathcal{E}(G_n^{\ddot{Y}})} k_{\mathfrak{X}}(X_i^\omega, X_j^\omega) - \frac{1}{n}\sum_{i=1}^n \frac{1}{d_i^Z} \sum_{j:(i,j)\in\mathcal{E}(G_n^Z)} k_{\mathfrak{X}}(X_i^\omega, X_j^\omega)}{\frac{1}{n}\sum_{i=1}^n k_{\mathfrak{X}}(X_i^\omega, X_j^\omega) - \frac{1}{n}\sum_{i=1}^n \frac{1}{d_i^Z} \sum_{j:(i,j)\in\mathcal{E}(G_n^Z)} k_{\mathfrak{X}}(X_i^\omega, X_j^\omega)} \tag{10}$$

where $\theta_{KPC} = (k_{\mathfrak{X}}, k)$ with the kernel $k_{\mathfrak{X}}$ and its parameters as well as the number of nearest-neighbors k.

While we could interchange X^ω and Y in the KPC owing to the symmetry of CI (Dawid, 1979), and this can lead to different results since conditional dependence measures are

not necessarily symmetric, our simulations indicate that the proposed order gives superior performance. This outcome may be attributed to the improved performance of the geometric graphs on the lower dimensional space $\mathcal{Y} \times \mathcal{Z}$ to the graphs for (X^ω, Z) on the higher dimensional $\mathfrak{X} \times \mathcal{Z}$.

The CPT-KPC-based CIT (CPT-KPC) depends on $\theta_{CPT,KPC} = (\theta_{CPT}, \theta_{KPC})$ where θ_{CPT} consists of the choice and the parameters of the learning model approximating $\mathbb{P}^{Y|Z}$ and the number M of constructed approximate CI samples in the CPT, and θ_{KPC} consists of the choice of the kernel $k_{\mathfrak{X}}$ and its parameters, and the number of nearest neighbors in the knn graphs. The KPC accounts for the dimension of the feature representations by embedding them through a flexibly chosen kernel. In addition, it would even allow for non-Euclidean feature representations such as shapes by the use of a suitable kernel function on $\mathfrak{X} \times \mathfrak{X}$, as well as non-scalar responses Y and confounders Z as long as they are elements of metric spaces. Nevertheless, the kernel $k_{\mathfrak{X}}$ affects the graph KPC, which is a common problem for kernel-based conditional dependence measures, see for example Muandet et al. (2017), and thus, can decrease the power of the CPT-KPC. To study this dependence on the kernel, we analyzed the results for several kernels in a smaller simulation study based on the implementation of the KPC[3], and selected the Gaussian kernel since it performed best over multiple settings. In addition, the geometric graph functionals defined on $\mathcal{Z}$ and $\mathcal{Y} \times \mathcal{Z}$ measure the proximity within the confounder as well as the response. However, the geometric graph functionals depend on the number of nearest neighbors selected, where 1-nn leads to less biased estimates of the KPC while a larger number of neighbors is recommended to increase the power if the KPC is used as test statistic. For example, Lin and Han (2023) examined the effect of the number of nearest neighbors on the power of unconditional independence tests, and Huang et al. (2022) recommend in their implementation to select $k \approx 0.05n$ for samples smaller than 1000 and then increase k sublinearly in n for their variable selection algorithm. Consequently, in the smaller study, we examined `Deep-CPT-KPC`s for varying numbers of nearest neighbors and selected $k = 10$ to trade-off between T1E control and power. Then, we evaluate the `Deep-CPT-KPC`s together with Euclidean knn graphs, since this was the recommended choice in Huang et al. (2022, Remark 7). This leads to a time complexity of $\mathcal{O}(kn \log n)$ for estimating the graph KPC in (10), see Huang et al. (2022). Together with the time complexity of the CPT consisting of the generalized additive model, this leads to $\mathcal{O}((M+1)kn \log n + ns^2)$ for CPT-KPCs. In summary, the hypothesis tested by the CPT-KPC is only constrained compared to (2) through weak, in the kernel literature standard, assumptions on the RKHS $\mathcal{H}_{\mathfrak{X}}$ and the kernel $k_{\mathfrak{X}}$ (Huang et al., 2022, Remark 2), or $\mathcal{H}_{\mathcal{Y}}$ and the kernel $k_{\mathcal{Y}}$ if we interchange X^ω and Y. This allows for T1E control over a large $\mathcal{P}_0^\omega$ and correspondingly large $\mathcal{P}_0$. Moreover, the guaranteed T1E excess bound of the CPT together with the consistency of the KPC seems promising in terms of T1E control as well as power of the CPT-KPC. Nevertheless, since either the kernel embedding of X^ω or Y is targeted, the test will detect small changes in $X^\omega|Y,Z$ compared to $X^\omega|Z$, or $Y|X^\omega, Z$ compared to $Y|Z$, respectively, better. We call DNCITs based on the CPT together with the KPC `Deep-CPT-KPC`.

3. The KPC package can be found under `https://github.com/zh2395/KPC`.

D.2 The Kernel Conditional Independence Test and the Randomized Correlation Test

We describe briefly the kernel-based RCoT from Strobl et al. (2019) and its approximation of the KCIT from Zhang et al. (2011). The RCoT approximates the hypothesis and test statistic of the KCIT through

$$H^{\omega}_{0,RCoT} : \|\mathcal{C}_{AB|C}\|^2_F = 0 \quad vs. \quad H^{\omega}_{1,RCoT} : \|\mathcal{C}_{AB|C}\|^2_F > 0$$

estimated by

$$T^{\omega}_{\theta_{RCoT}}((X^{\omega}, Y, Z)^n) = n\text{tr}(\widehat{\mathcal{C}}_{AB|C}\widehat{\mathcal{C}}^{\top}_{AB|C})$$

where

$$\begin{aligned}
\mathcal{C}_{AB|C} &= \mathbb{E}[(A_i - \mathbb{E}[A|C])(B_i - \mathbb{E}[B|C])^{\top}] \\
\widehat{\mathcal{C}}_{AB|C} &= \frac{1}{n-1}\sum_{i=1}^{n}[(A_i - \widehat{\mathbb{E}}(A|C))(B_i - \widehat{\mathbb{E}}(B|C))^{\top}] \\
A_i &= \phi^{X^{\omega}}(X^{\omega}_i) = \{\phi^{X^{\omega}}_1(X^{\omega}_i), \ldots, \phi^{X^{\omega}}_a(X^{\omega}_i)\}, \quad \phi^{X^{\omega}}_j(X^{\omega}_i) \in \mathcal{G}_{X^{\omega}}, \forall j, \\
B_i &= \phi^{Y}(Y) = \{\phi^{Y}_1(Y_i), \ldots, \phi^{Y}_b(Y_i)\}, \quad \phi^{Y}_k(Y_i) \in \mathcal{G}_Y, \forall k, \\
C_i &= \phi^{Z}(Z_i) = \{\phi^{Z}_1(Z_i), \ldots, \phi^{Z}_c(Z_i)\}, \quad \phi^{Z}_l(Z_i) \in \mathcal{G}_Z, \forall l,
\end{aligned}$$

$\mathcal{G}_{X^{\omega}}, \mathcal{G}_Y, \mathcal{G}_Z$ are spaces set to be the support of the process $\sqrt{2}\cos(W^{\top}\cdot + B), W \sim \mathbb{P}^W, B \sim Unif([0, 2\pi])$ with $\mathbb{P}^W$ set to a Gaussian distribution with standard deviations $\sqrt{\sigma^2_A/2}$, $\sqrt{\sigma_B/2}$, and $\sqrt{\sigma_C/2}$, and $\|\cdot\|_F$ denotes the Frobenius norm. That means, a, b and c functions are drawn from the spaces $\mathcal{G}_{X^{\omega}}, \mathcal{G}_Y$ and $\mathcal{G}_Z$, respectively. Moreover, the conditional expectations $\mathbb{E}(A|C)$ and $\mathbb{E}(B|C)$ are approximated with linear ridge regression solutions. Thus, $\theta_{RCoT} = (a, b, c, \sigma_A, \sigma_B, \sigma_C, \lambda)$ with a, b, c for the number of Fourier functions of A, B, C, the smoothing parameters $\sigma_A, \sigma_B, \sigma_C$ of the radial basis function kernels and the regularization parameter λ of the linear ridge regressions.

This approximates KCIT's hypotheses and the corresponding test statistic in three ways: 1.) Replacing $\ddot{X}^{\omega} = (X^{\omega}, Z)$ by X^{ω}. This corresponds to a test of weak CI of the data mapped into RKHSs instead of strong CI, as discussed in Li and Fan (2020). Nevertheless, RCoT is usually similar in its performance compared to the corresponding Randomized Conditional Independence Test (RCIT) using $\ddot{X}^{\omega}$, which was also developed in Strobl et al. (2019); 2.) The kernel maps of $\ddot{X}^{\omega}$ and Y are replaced by low dimensional, random Fourier features. This is based on the approximation of continuous shift-invariant kernels through low-dimensional Fourier features, which was shown to perform well on large-scale data sets (Rahimi and Recht, 2007); 3.) The conditional expectations $\mathbb{E}[A|Z]$ and $\mathbb{E}[B|Z]$ are approximated by linear ridge regressions on C, which implies an approximation of the empirical partial cross-covariance matrix $\widehat{\mathcal{C}}_{AB|Z}$ through $\widehat{\mathcal{C}}_{AB|C}$. It was shown in Sutherland and Schneider (2015) that, under appropriate assumptions, the linear ridge regression solutions converge with an exponential rate in c to the corresponding kernel ridge regression, justifying this approximation.

Asymptotic distributions and corresponding approximations are derived for the test statistics of KCIT and RCoT. The RCoT depends on the hyperparameters a,b,c,σ_A,σ_B,σ_C,λ. As described in Zhang et al. (2011); Strobl et al. (2019), $\lambda \approx 0.1$ is a reasonable choice for low-dimensional Z as in our setting. Moreover, $\sigma_A, \sigma_B, \sigma_C$ are chosen empirically on the first 500 observations or in sample, if the number of observations is smaller than 500. Compared to other nonparametric CITs discussed later, these are relatively few hyperparameters and the test performed well in our simulations and the slightly adapted default parameters $a = 5, b = 5, c = 200$ for two confounders. For more than two confounders, the ridge regressions have to be additionally adapted to control T1E. The RCoT is faster than the KCIT due to the three approximations discussed previously, as well as an approximation of the asymptotic distribution under the null of $\widehat{\mathcal{C}}_{AB|C}$ by a Lindsay-Pilla-Basak approximation, which makes a comparison with the null hypothesis computationally inexpensive. In particular, the RCoT has a time complexity of $\mathcal{O}(((\dim_{X^\omega} + \dim_Z)a + (\dim_Y + \dim_Z)b + \dim_Z c + (a+b)c^2)n)$, which reduces to $\mathcal{O}(n)$ for fixed $a, b, c, \dim_{X^\omega}, \dim_Y, \dim_Z$ (Strobl et al., 2019, Proposition 6). Besides that, the RCoT covers a wide range of functional relationships between the random objects due to their Fourier transformations. Additionally, the projection of X^ω into the lower dimensional space $\mathcal{G}_{X^\omega}$ reduces the dimension of the feature representations X^ω, which is particularly advantageous in our setting and allows for relatively high-dimensional X^ω. Therefore, in contrast to the recommendation in Strobl et al. (2019), we typically aim for a larger number of Fourier features a to capture most information in the high dimension of X^ω. Furthermore, the test is stably implemented in R[4]. Moreover, we note that for small sample sizes, it may be preferable to use the KCIT due to a failure of the RCoT in the T1E control in these settings due to the approximations, see for example the simulation studies also in Sen et al. (2017); Zhang et al. (2023a). Furthermore, for several confounders, the ridge regressions would have to be adapted or other CITs may be more appropriate, as discussed in Sen et al. (2017). In summary, the hypotheses considered in the RCoT are restricted compared to the hypotheses (2) only by the relatively weak assumption on the kernel $k_{\tilde{\mathcal{X}}} k_{\mathcal{Y}}$ and the RKHS $\mathcal{H}_{\mathcal{Z}}$, as well as the restriction to approximate Gaussian kernels in $\mathbb{P}^{\hat{W}}$ and the corresponding, well-founded approximations of the KCIT. We call the corresponding DNCIT `Deep-RCoT`.

D.3 Prediction-Based Conditional Independence Tests

We evaluate the applicability of prediction-based CITs to vector-scalar-valued data. We can apply them by testing for an increase in the prediction accuracy of Y using X^ω, Z compared to the prediction solely with Z, removing the predictive information within X^ω. This can be done with various approaches as reviewed in Covert et al. (2021) from the perspective of model explanations, and translates the CIT typically to a weaker mean dependence test, accounting only for parts of the conditional distribution of Y encoded through the loss function used to fit the prediction model. In the following, we describe the approaches developed explicitly for CI testing, namely the CPI (Watson and Wright, 2021), the FCIT (Chalupka et al., 2018) and the PCIT (Burkart and Király, 2017).

Therefore, let $f \in \mathcal{M}_{\mathfrak{X}\mathcal{Z}\to\mathcal{Y}}, \mathcal{M}_{\mathfrak{X}\mathcal{Z}\to\mathcal{Y}} : \mathfrak{X} \times \mathcal{Z} \to \mathcal{Y}$ be a prediction function for the scalar Y using the feature representations X^ω and the confounders Z. Then, the CPI

4. The RCIT package can be found under `https://github.com/ericstrobl/RCIT`.

removes the predictive information of X^ω on Y by constructing knockoffs $\widetilde{X^\omega}$, which were introduced by Barber and Candès (2015); Candes et al. (2018). Particularly, for the risk $R_L(f,(X^\omega,Y,Z)) = \mathbb{E}[L(f,(X^\omega,Y,Z))]$ and a non-negative, real-valued loss function $L : \mathcal{M}_{\mathfrak{X}\mathcal{Z}\to\mathcal{Y}} \times (\mathfrak{X}\times\mathcal{Y}\times\mathcal{Z}) \to \mathbb{R}_{\geq 0}$, the CPI is a measure of conditional dependence as

$$CPI((X^\omega,Y,Z)) = \mathbb{E}[L(f,(\widetilde{X^\omega},Y,Z)) - L(f,(X^\omega,Y,Z))]$$

Then, they propose to test (2) by testing

$$\begin{aligned} & H^\omega_{0,CPI} : CPI((X^\omega,Y,Z)) \leq 0 \iff R_L(f,(\widetilde{X^\omega},Y,Z)) \leq R_L(f,(X^\omega,Y,Z)) \\ vs. \quad & H^\omega_{1,CPI} : CPI((X^\omega,Y,Z)) > 0 \iff R_L(f,(\widetilde{X^\omega},Y,Z)) > R_L(f,(X^\omega,Y,Z)). \end{aligned}$$

The CPI can be estimated by the empirical risks leading to the corresponding test statistic

$$T^\omega_{\theta_{CPI}}((X^\omega,Y,Z)^n) = \frac{1}{n_{test}}\sum_{i=1}^{n_{test}} L(f,(\widetilde{X^\omega}_i,Y_i,Z_i)) - \frac{1}{n_{test}}\sum_{i=1}^{n_{test}} L(f,(X^\omega_i,Y_i,Z_i))$$

over a test data set $(X^\omega,Y,Z)^{n_{test}}$ of size n_{test}, here obtained after splitting $(X^\omega,Y,Z)^n$ into train and test data set. Parameters θ_{CPI} consist of the knockoff construction parameters, the loss L and the prediction model f as well as its parameters.

In contrast to the CPI and its use of knockoffs, the PCIT and FCIT remove X^ω from the prediction to obtain a prediction function $g \in \mathcal{M}_{\mathcal{Z}\to\mathcal{Y}}, \mathcal{M}_{\mathcal{Z}\to\mathcal{Y}} : \mathcal{Z}\to\mathcal{Y}$, by regressing Y only on the confounder Z. Then, the authors propose to test (2) by testing

$$\begin{aligned} & H^\omega_{0,FCIT/PCIT} : R_L(g,(Y,Z)) \leq R_L(f,(X^\omega,Y,Z)) \\ vs. \quad & H^\omega_{1,FCIT/PCIT} : R_L(g,(Y,Z)) > R_L(f,(X^\omega,Y,Z)), \end{aligned}$$

which can again be estimated by the corresponding empirical risks. While the FCIT implements a decision tree algorithm to learn g and f, the PCIT allows fitting a set of regression and classification models as well as their ensembles. However, PCIT's implementation is currently unstable and not further used in our study.

For the comparison to the null hypothesis, any paired mean difference test such as the paired t-test or Fisher exact test can be used to obtain p-values. All three tests can rely on asymptotic normality results of the empirical risk estimates.

We call the corresponding DNCITs `Deep-CPI` and `Deep-FCIT`. They reduce conditional independence to conditional variable importance. Thereby CI testing is translated to mean dependence or at least loss dependent CI testing, which has the advantage that it draws on a large literature of variable importance testing, while often being sufficient to detect deviations from the null. However, this means that the tests can only have power for data generating processes for which the difference between $\mathbb{P}^{Y|X^\omega,Z}$ and $\mathbb{P}^{Y|\widetilde{X^\omega},Z}$ or $\mathbb{P}^{Y|Z}$, respectively, is detectable by the chosen loss. Particularly, the tests focus on specific properties of the distribution encoded in the loss L instead of the whole conditional distribution of Y, resulting in a loss of power for differences in other properties of the distribution. Additionally, it usually restricts the problem to a specific learner or an ensemble of learners and the conditional variable importance for that learner(s). Therefore, while the tests themselves

do not impose restrictions w.r.t. X^ω, Y, Z and their functional relationships, the tests impose these implicitly through the used learners. In particular, prediction-based DNCITs only account for functional relationships between X^ω and Y which can be approximated by the learner. This leads to a potential loss in power if f does not cover the true functional relationship between X^ω and Y, and an increase in the T1E if g does not estimate the true functional relationship between Y and Z well enough or if the knockoffs are not well constructed, similar to the T1E excess for CPT-based CITs. On the one hand this flexibility in the choice of the learner allows for models adapted to a specific data setting. On the other hand, it is difficult to obtain a default CIT, which would ensure consistency over test results, making these tests rather hyperparameter dependent. Moreover, the sample split necessary to fit the learner results in a reduced power. Finally, an advantage can be that their implementations[5] are based on often stable implementations of learners for the functions f and g. Particularly, the CPI-based CIT has a time complexity of $\mathcal{O}(ue + v + w)$ where u is the complexity of the learner f, e of the empirical risk estimator, v of the knockoff sampler and w of the inference procedure such as the t-test (Watson and Wright, 2021). The FCIT has a time complexity, without the inclusion of the knockoff but with additional costs to fit the learner g, of $\mathcal{O}(n^2 \log n(p+q))$ (Chalupka et al., 2018).

As a final remark on prediction-based CITs, we note that they could be applied directly to the images, either by removing X in a model refit as in FCIT and PCIT, or by using knockoffs for X. However, this exacerbates the challenges described in the previous paragraph. That is, T1E, power, and complexity are affected as follows: First, the T1E increases for misconstruction of the knockoffs of X, which are harder to construct than knockoffs of X^ω. Second, the power decreases when the model cannot account for an increase in predictive accuracy using X and Z compared to the predictive accuracy using Z alone. While there exist many models to estimate the additional predictive accuracy using X^ω and Z, often working well with established automated hyperparameter tuning, the models such as neural networks for X and Z are often strongly dependent on their hyperparameters without simple tuning methods, making the performance of the tests even more hyperparameter dependent and difficult to fit on small data sets. Finally, the models for learning f increase in complexity, which greatly increases the time complexity for each test. This makes the corresponding tests computationally expensive or even prohibitive, especially for multiple tests including the images, compared to DNCITs using the same feature representations over all multiple tests. Therefore, we do not investigate this approach further in this paper.

D.4 The Projected Covariance Measure

The PCM (Lundborg et al., 2024) was developed as a conditional mean dependence test and tests

$$H^\omega_{0,PCM} : \mathbb{E}[Y|X^\omega, Z] = \mathbb{E}[Y|Z] \quad vs. \quad H^\omega_{1,PCM} : \mathbb{E}[Y|X^\omega, Z] \neq \mathbb{E}[Y|Z].$$

5. The FCIT package is available in python under `https://github.com/kjchalup/fcit`, the PCIT package under `https://github.com/alan-turing-institute/pcit`, and the CPI package in R under `https://github.com/bips-hb/cpi`.

As a test statistic, the authors propose

$$T^{\omega}_{\theta_{PCM}} = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} L_i}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} L_i^2 - (\frac{1}{n} \sum_{i=1}^{n} L_i)^2}}$$

where $L_i = \{Y_i - \hat{m}(Z_i)\}\{\hat{f}(X_i^{\omega}, Z_i) - \hat{m}_{\hat{f}}(Z_i)\}$ with an estimate $\hat{f}(X^{\omega}, Z)$ for $f(X^{\omega}, Z) = \frac{\mathbb{E}[Y|X^{\omega},Z] - \mathbb{E}[Y|Z]}{Var(Y|X^{\omega},Z)}$, and estimates $\hat{m}(Z)$ and $\hat{m}_{\hat{f}}(Z)$ for $\mathbb{E}[Y|Z]$ and $\mathbb{E}[\hat{f}(X^{\omega}, Z)|Z, \hat{f}]$, respectively. To avoid inflated T1Es due to the estimation of f, the PCM relies on sample splitting. To use the data more efficiently than splitting it into two halves, the p-value can be computed by averaging the p-values of the PCMs applied to several folds. This should increase the power of the test, as discussed in Kook and Lundborg (2024), but also increases its runtime.

Under $H^{\omega}_{0,PCM}$, the test statistic converges to a standard normal distribution, given assumptions that rely most notably on the performance of the regression models, $\hat{m}$ and $\hat{m}_{\hat{f}}$, and on the estimation of f. Specifically, the test controls the T1E if the product of the mean squared prediction errors (MSPEs) of $\hat{m}$ and $\hat{m}_{\hat{f}}$ converges at a rate of n^{-1}. Then, a one-sided test for large values can be applied, since the test statistic is positive under the alternative hypothesis.

Compared to (2), PCM's null, $H^{\omega}_{0,PCM}$, only tests whether $\mathbb{E}[Cov(Y, f(X^{\omega}, Z)|Z)] = 0$. Thus, the PCM is powerless if $Y \not\perp\!\!\!\perp X^{\omega}|Z$ but $\mathbb{E}[Cov(Y, f(X^{\omega}, Z)|Z)] = 0$ holds. Further details on this comparison between CI and conditional mean dependence are also given in Lundborg et al. (2024). Apart from this potential loss of power when considering the conditional mean of Y instead of its conditional distribution, the PCM is well suited to test (2) since it includes the high-dimensional X^{ω} and Z as inputs of regression models in the test statistic. Thereby, the PCM avoids having X^{ω} as output as in other residual-based CITs like the GCM, making it more suitable for feature representations. Furthermore, the PCM performs well compared to other conditional mean dependence tests (Williamson et al., 2023; Dai et al., 2022; Cai et al., 2025).

The PCM depends on the parameters $\theta_{PCM} = (\theta_{\hat{f}}, \theta_{\hat{m}}, \theta_{\hat{m}_{\hat{f}}}, K)$, where $\theta_{\hat{f}}, \theta_{\hat{m}}, \theta_{\hat{m}_{\hat{f}}}$ are the parameters of the regression models used to estimate f, and the regression models $\hat{m}$ and $\hat{m}_{\hat{f}}$, respectively, and K is the number of folds used to split the sample. Although the test could be applied directly to structured objects, such as images, the remarks at the end of Appendix D.3 apply to the PCM as well. Specifically, there are no theoretical guarantees that more complex models fitted to structured objects will converge fast enough to satisfy the MSPE assumptions. Furthermore, current packages do not directly implement such models, focusing instead on tabular regression models (Kook and Lundborg, 2024). Nevertheless, this implementation works well when applied to feature representations of structured objects, as in the DNCIT, instead of the structured objects directly. Thus, we applied the PCM with $K = 1$ and its default random forest regressions (Wright and Ziegler, 2017) to estimate all quantities in f as well as $\hat{m}$ and $\hat{m}_{\hat{f}}$. The time complexity of the PCM depends on the repetition of PCMs K and the regression models used. For the random forest regressions, the PCM has time complexity of $\mathcal{O}((p+q)n \log ntree_n)$ for the training, where $p+q$ is the feature representation dimension plus the confounder dimension and $tree_n$ is the number of trees, as well as $\mathcal{O}(tree_{depth}(p+q))$ for predicting where $tree_{depth}$ is the tree

depth, together with a cross-validation with k being the number of folds for each regression model. In summary, the PCM is well suited to test (2) due to the inclusion of X^ω as an input to regression models, its theoretical guarantees based on achievable assumptions, and its stable implementation. We call the corresponding DNCIT `Deep-PCM`. For future research, the PCM could be extended to the conditional distribution of Y to also have power in settings where $Y \not\perp\!\!\!\perp X^\omega|Z$ but $\mathbb{E}[Cov(Y, f(X^\omega, Z)|Z)] = 0$. Furthermore, it would be interesting to study empirically whether X can be used directly as input. Finally, an implementation for categorical Y is currently lacking.

D.5 The Conditional Mutual Information K-Nearest-Neighbor Test

The metric-based CMIknn CIT of Runge (2018) is based on the CMI as a test statistic, testing

$$H^\omega_{0,CMIknn} : CMI((X^\omega, Y, Z)) = \int\int\int dxdydzp(x, y, z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)} = 0$$
$$vs. \quad H^\omega_{1,CMIknn} : CMI((X^\omega, Y, Z)) > 0,$$

since $CMI((X^\omega, Y, Z)) = 0$ if and only if $X^\omega \perp\!\!\!\perp Y|Z$ (Runge, 2018). Then, the authors present the estimator

$$T^{CMIknn}((X^\omega, Y, Z)^n) = \psi(k_{CMI}) + \frac{1}{n}\sum_{i=1}^{n} \left[\psi(k_i^z) - \psi(k_i^{xz}) - \psi(k_i^{yz})\right],$$

where $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \ln \Gamma(x)$, $\Gamma(\cdot)$ the Gamma function, k_{CMI} is a pre-specified number of nearest neighbors in $\mathfrak{X} \times \mathcal{Y} \times \mathcal{Z}$ defining implicitly the local distances $\epsilon_i, i = 1, \ldots, n$ and $k_i^z, k_i^{xz}, k_i^{yz}$ are the corresponding numbers of nearest neighbors for $i = 1, \ldots, n$ in $\mathcal{Z}, \mathfrak{X} \times \mathcal{Z}$ and $\mathcal{Y} \times \mathcal{Z}$, respectively, with distance strictly smaller than ϵ_i.

This test statistic is compared to the null hypothesis through a local permutation scheme. In particular, the k_{perm}-nearest neighbors in $\mathcal{Z}$ are computed and for each $Y_i, i = 1, \ldots, n$ its neighbors are selected by these nearest neighbors in $\mathcal{Z}$. Then, each Y_i in Y^n is locally, randomly permuted M times such that each Y_i is, if possible, only used once in the permuted $(Y^{(m)})^n, m = 1, \ldots, M$. Finally, the p-value is obtained through $p^\omega = M^{-1} \sum_{m=1}^{M} \mathbb{1}\{\widehat{T}^{CMIknn}((X^\omega, Y^{(m)}, Z)^n) \geq \widehat{T}^{CMIknn}((X^\omega, Y, Z)^n)\}$.

The CMIknn CIT[6] is applicable to arbitrary dimensions of X^ω, Y, Z, which makes it applicable to the feature representations of the images and even more generally, metric spaces $\mathfrak{X}, \mathcal{Y}, \mathcal{Z}$. It depends only on the hyperparameters k_{CMI} and k_{perm}, where the authors suggest a low, fixed $k_{perm} \in \{5, \ldots, 10\}$ and a rule-of-thumb of $k_{CMI} \approx 0.1n$ or $0.2n$. The T1E and power increase for larger k_{perm}, while it is theoretically guaranteed only for $k_{perm} = 1$ that $(X^\omega, Y^{(m)}, Z)^n, m = 1, \ldots, M$ are approximately sampled from $\mathcal{P}_0^{\mathfrak{X}}$ (Sen et al., 2017). The power of the test increases in k_{CMI} up to some maximum in the existing simulation studies in Runge (2018). The test has a computational cost of $\mathcal{O}(n^2)$ for the search for nearest neighbors in $\hat{T}^{CMIknn}$ and the permutation scheme, and scales roughly linearly in k_{CMI} and $(p + q + 1)$, where k_{CMI} is the number of nearest neighbors used to

6. The CMIknn package can be found under `https://github.com/jakobrunge/tigramite`.

estimate T^{CMIknn} and $p+q+1$ is the dimension of (X^ω, Y, Z) (Runge, 2018). We call the corresponding DNCIT `Deep-CMIknn`.

References

A. Abrol, Z. Fu, M. Salman, R. Silva, Y. Du, S. Plis, and V. Calhoun. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nature communications*, 12(1):353, 2021.

F. Alfaro-Almagro, M. Jenkinson, N. K. Bangerter, J. L. Andersson, L. Griffanti, G. Douaud, S. N. Sotiropoulos, S. Jbabdi, M. Hernandez-Fernandez, E. Vallee, et al. Image processing and quality control for the first 10,000 brain imaging datasets from UK biobank. *Neuroimage*, 166:400–424, 2018.

F. Alfaro-Almagro, P. McCarthy, S. Afyouni, J. L. Andersson, M. Bastiani, K. L. Miller, T. E. Nichols, and S. M. Smith. Confound modelling in UK biobank brain imaging. *NeuroImage*, 224:117002, 2021.

J. Ascorbebeitia, E. Ferreira, and S. Orbe. Testing conditional multivariate rank correlations: the effect of institutional quality on factors influencing competitiveness. *Test*, 31 (4):931–949, 2022.

R. Avinun, S. Israel, A. R. Knodt, and A. R. Hariri. Little evidence for associations between the big five personality traits and variability in brain gray or white matter. *NeuroImage*, 220:117092, 2020.

M. Azadkia and S. Chatterjee. A simple measure of conditional dependence. *The Annals of Statistics*, 49(6):3070–3102, 2021.

R. Barber and E. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

A. Bellot and M. van der Schaar. Conditional independence testing using generative adversarial networks. *Advances in Neural Information Processing Systems*, 32:2202–2211, 2019.

T. Berrett, Y. Wang, R. Barber, and R. Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):175–197, 2020.

V. I. Bogachev and M. A. S. Ruas. *Measure theory*, volume 2. Springer, 2007.

S. Burkart and F. Király. Predictive independence testing, predictive conditional independence testing, and predictive graphical modelling. *arXiv preprint arXiv:1711.05869*, 2017.

L. Cai, X. Guo, and W. Zhong. Test and measure for partial mean dependence based on machine learning methods. *Journal of the American Statistical Association*, 120(550): 833–845, 2025.

E. Candes, Y. Fan, L. Janson, and J. Lv. Panning for gold: 'Model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):551–577, 2018.

M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

K. Chalupka, P. Perona, and F. Eberhardt. Fast conditional independence test for vector variables with large sample sizes. *arXiv preprint arXiv:1804.02747*, 2018.

S. Chen, K. Ma, and Y. Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

I. Covert, S. Lundberg, and S.-I. Lee. Explaining by removing: A unified framework for model explanation. *The Journal of Machine Learning Research*, 22(1):9477–9566, 2021.

A. D. Dahlén, M. Miguet, H. B. Schiöth, and G. Rukh. The influence of personality on the risk of myocardial infarction in UK biobank cohort. *Scientific Reports*, 12(1):6706, 2022.

B. Dai, X. Shen, and W. Pan. Significance tests of feature relevance for a black-box learner. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15, 1979.

M. J. de Ruijter, A. Dahlén, G. Rukh, and H. B. Schiöth. Association of diligence and sociability with stroke: a UK biobank study on personality proxies. *Frontiers in Bioscience-Landmark*, 27(8), 2022.

N. Deb, P. Ghosal, and B. Sen. Measuring association on topological spaces using kernels and geometric graphs. *arXiv preprint arXiv:2010.01768*, 2020.

C. G. DeYoung, R. E. Beaty, E. Genç, R. D. Latzman, L. Passamonti, M. N. Servaas, A. J. Shackman, L. D. Smillie, R. N. Spreng, E. Viding, et al. Personality neuroscience: An emerging field with bright prospects. *Personality science*, 3, 2022.

J. M. Digman. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440, 1990.

G. Doran, K. Muandet, K. Zhang, and B. Schölkopf. A permutation-based kernel conditional independence test. In *Uncertainty in Artificial Intelligence (UAI 2014)*, pages 132–141. AUAI Press, 2014.

B. Duong and T. Nguyen. Conditional independence testing via latent representation learning. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 121–130. IEEE, 2022.

B. Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

A. Fisher, C. Rudin, and F. Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *The Journal of Machine Learning Research*, 20(177):1–81, 2019.

S. Genon, S. B. Eickhoff, and S. Kharabian. Linking interindividual variability in brain structure to behaviour. *Nature Reviews Neuroscience*, 23(5):307–318, 2022.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

M. Grosse-Wentrup, D. Janzing, M. Siegel, and B. Schölkopf. Identification of causal relations in neuroimaging data with latent confounders: An instrumental variable approach. *NeuroImage*, 125:825–833, 2016.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

L. Henschel, S. Conjeti, S. Estrada, K. Diers, B. Fischl, and M. Reuter. Fastsurfer-a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219:117012, 2020.

M. Huang, Y. Sun, and H. White. A flexible nonparametric test for conditional independence. *Econometric Theory*, 32(6):1434–1482, 2016.

Z. Huang, N. Deb, and B. Sen. Kernel partial correlation coefficient—a measure of conditional dependence. *The Journal of Machine Learning Research*, 23(1):9699–9756, 2022.

C. S. Hyatt, M. M. Owens, M. L. Crowe, N. T. Carter, D. R. Lynam, and J. D. Miller. The quandary of covarying: A brief review and empirical examination of covariate use in structural neuroimaging studies on psychological variables. *NeuroImage*, 205:116225, 2020.

O. P. John and S. Srivastava. The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin and O. P. John, editors, *Handbook of Personality: Theory and Research*, pages 102–138. Guilford Press, New York, 2 edition, 1999.

E. Katsevich and A. Ramdas. On the power of conditional independence testing under model-x. *Electronic Journal of Statistics*, 16(2):6348–6394, 2022.

S. Kharabian Masouleh, S. B. Eickhoff, F. Hoffstaedter, S. Genon, and A. D. N. Initiative. Empirical examination of the replicability of associations between brain structure and psychological variables. *elife*, 8:e43464, 2019.

I. Kim, M. Neykov, S. Balakrishnan, and L. Wasserman. Local permutation tests for conditional independence. *The Annals of Statistics*, 50(6):3388–3414, 2022.

D. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

M. Kirchler, S. Khorasani, M. Kloft, and C. Lippert. Two-sample testing using deep learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1387–1398. PMLR, 2020.

M. Kirchler, S. Konigorski, M. Norden, C. Meltendorf, M. Kloft, C. Schurmann, and C. Lippert. transferGWAS: GWAS of images using deep transfer learning. *Bioinformatics*, 38 (14):3621–3628, 2022.

L. Kook and A. R. Lundborg. Algorithm-agnostic significance testing in supervised learning with multimodal data. *Briefings in Bioinformatics*, 25(6):bbae475, 2024.

E. Lehmann, J. Romano, and G. Casella. *Testing Statistical Hypotheses*. Springer Nature Switzerland AG 2022, fourth edition, 2022.

C. Li and X. Fan. On nonparametric conditional independence tests for continuous variables. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(3):e1489, 2020.

S. Li, M. Sesia, Y. Romano, E. Candès, and C. Sabatti. Searching for robust associations with a multi-environment knockoff filter. *Biometrika*, 109(3):611–629, 2022.

Z. Lin and F. Han. On boosting the power of chatterjee's rank correlation. *Biometrika*, 110 (2):283–299, 2023.

F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning*, pages 6316–6326. PMLR, 2020.

I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

A. R. Lundborg, R. D. Shah, and J. Peters. Conditional independence testing in hilbert spaces with applications to functional data analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1821–1850, 2022.

A. R. Lundborg, I. Kim, R. D. Shah, and R. J. Samworth. The projected covariance measure for assumption-lean variable significance testing. *The Annals of Statistics*, 52 (6):2851–2878, 2024.

S. Marek, B. Tervo-Clemmens, F. J. Calabro, D. F. Montez, B. P. Kay, A. S. Hatoum, M. R. Donohue, W. Foran, R. L. Miller, T. J. Hendrickson, et al. Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902):654–660, 2022.

B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

K. L. Miller, F. Alfaro-Almagro, N. K. Bangerter, D. L. Thomas, E. Yacoub, J. Xu, A. J. Bartsch, S. Jbabdi, S. N. Sotiropoulos, J. L. Andersson, et al. Multimodal population brain imaging in the UK biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523–1536, 2016.

T. P. Morris, I. R. White, and M. J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11):2074–2102, 2019.

K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15(4):869, 2005.

M. Neykov, S. Balakrishnan, and L. Wasserman. Minimax optimal conditional independence testing. *The Annals of Statistics*, 49(4):2151–2177, 2021.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

R. Pogodin, N. Deka, Y. Li, D. Sutherland, V. Veitch, and A. Gretton. Efficient conditionally invariant representation learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=dJruFeSRym1`.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20:1177–1184, 2007.

A. Rakowski, R. Monti, and C. Lippert. TransferGWAS of T1-weighted brain MRI data from UK Biobank. *PLoS Genetics*, 20(12):e1011332, 2024.

K. Raza and N. Singh. A tour of unsupervised deep learning for medical image analysis. *Current Medical Imaging*, 17(9):1059–1077, 2021.

J. Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 938–947. PMLR, 2018.

J. Runge, A. Gerhardus, G. Varando, V. Eyring, and G. Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, 2023.

A. Salehi, S. Khan, G. Gupta, B. Alabduallah, A. Almjally, H. Alsolai, T. Siddiqui, and A. Mellit. A study of cnn and transfer learning in medical imaging: Advantages, challenges, future scope. *Sustainability*, 15(7):5930, 2023.

M. Scetbon, L. Meunier, and Y. Romano. An asymptotic test for conditional independence using analytic kernel embeddings. In *International Conference on Machine Learning*, pages 19328–19346. PMLR, 2022.

C. Scheidegger, J. Hörrmann, and P. Bühlmann. The weighted generalised covariance measure. *Journal of Machine Learning Research*, 23(273):1–68, 2022.

S. Sekimitsu, J. Wang, T. Elze, A. V. Segrè, J. L. Wiggs, and N. Zebardast. Interaction of background genetic risk, psychotropic medications, and primary angle closure glaucoma in the UK biobank. *Plos one*, 17(6):e0270530, 2022.

R. Sen, A. Suresh, K. Shanmugam, A. Dimakis, and S. Shakkottai. Model-powered conditional independence test. *Advances in neural information processing systems*, 30:2955–2965, 2017.

D. Servén and C. Brummitt. pygam: Generalized additive models in python, Mar. 2018. URL `https://doi.org/10.5281/zenodo.1208723`.

R. Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

C. Shi, T. Xu, W. Bergsma, and L. Li. Double generative adversarial networks for conditional independence testing. *The Journal of Machine Learning Research*, 22(1):13029–13060, 2021.

H. Shi, M. Drton, and F. Han. On azadkia–chatterjee's conditional dependence coefficient. *Bernoulli*, 30(2):851–877, 2024.

D. J. Smith, B. I. Nicholl, B. Cullen, D. Martin, Z. Ul-Haq, J. Evans, J. M. Gill, B. Roberts, J. Gallacher, D. Mackay, et al. Prevalence and characteristics of probable major depression and bipolar disorder within UK biobank: cross-sectional study of 172,751 participants. *PloS one*, 8(11):e75362, 2013.

S. Smith, F. Alfaro-Almagro, and K. Miller. UK biobank brain imaging documentation. `biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf`, 09 2022. Online; accessed 21.03.2024.

S. M. Smith, F. Alfaro-Almagro, and K. L. Miller. Uk biobank brain imaging documentation. Technical report, UK Biobank, 2024. URL `https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf`. Resource 1977.

T. Spisak. Statistical quantification of confounding bias in machine learning models. *GigaScience*, 11:giac082, 2022.

E. Strobl, K. Zhang, and S. Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1), 2019.

L. Su and H. White. A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*, 141(2):807–834, 2007.

L. Su and H. White. Testing conditional independence via empirical likelihood. *Journal of Econometrics*, 182(1):27–44, 2014.

C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, et al. UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3): e1001779, 2015.

D. Sutherland and J. Schneider. On the error of random Fourier features. In *Uncertainty in Artificial Intelligence (UAI 2015)*, pages 862–871, 2015.

D. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*, 2021.

J. Valverde, V. Imani, A. Abdollahzadeh, R. De Feo, R. Prakash, M.and Ciszek, and J. Tohka. Transfer learning in magnetic resonance brain imaging: A systematic review. *Journal of imaging*, 7(4):66, 2021.

M. van der Meulen, J. M. Amaya, O. M. Dekkers, and O. C. Meijer. Association between use of systemic and inhaled glucocorticoids and changes in brain volume and white matter microstructure: a cross-sectional study using data from the UK biobank. *BMJ open*, 12 (8):e062446, 2022.

S. Wang, B. Yuan, T. Tony Cai, and H. Li. Phylogenetic association analysis with conditional rank correlation. *Biometrika*, 111(3):881–902, 2024.

W. Wang and L. Janson. A high-dimensional power analysis of the conditional randomization test and knockoffs. *Biometrika*, 109(3):631–645, 2022.

X. Wang, W. Pan, W. Hu, Y. Tian, and H. Zhang. Conditional distance correlation. *Journal of the American Statistical Association*, 110(512):1726–1734, 2015.

D. Watson and M. Wright. Testing conditional independence in supervised learning algorithms. *Machine Learning*, 110(8):2107–2129, 2021.

B. D. Williamson, P. B. Gilbert, M. Carone, and N. Simon. Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1):9–22, 2021.

B. D. Williamson, P. B. Gilbert, N. R. Simon, and M. Carone. A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 118(543):1645–1658, 2023.

S. Wood. Package 'mgcv'. *R package version*, 1(29):729, 2015.

S. Wood. Inference and computation with generalized additive models and their extensions. *Test*, 29(2):307–339, 2020.

M. N. Wright and A. Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of statistical software*, 77:1–17, 2017.

T. Yarkoni. Neurobiological substrates of personality: A critical overview. In M. Mikulincer, P. R. Shaver, M. L. Cooper, and R. J. Larsen, editors, *APA Handbook of Personality and Social Psychology, Volume 4: Personality Processes and Individual Differences*, pages 61–83. American Psychological Association, Washington, DC, 2015. doi: 10.1037/14343-003.

H. Zhang, S. Zhou, K. Zhang, and J. Guan. Causal discovery using regression-based conditional independence tests. In *AAAI Conference on Artificial Intelligence*, volume 31(1), 2017.

H. Zhang, Y. Xia, K. Zhang, S. Zhou, and J. Guan. Conditional independence test based on residual similarity. *ACM Transactions on Knowledge Discovery from Data*, 17(8):1–18, 2023a.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Uncertainty in Artificial Intelligence (UAI 2011)*, pages 804–813. AUAI Press, 2011.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *International conference on machine learning*, pages 819–827. Pmlr, 2013.

Y. Zhang, M. Brady, and S. Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2002.

Y.-R. Zhang, Y.-T. Deng, Y.-Z. Li, R.-Q. Zhang, K. Kuo, Y.-J. Ge, B.-S. Wu, W. Zhang, A. D. Smith, J. Suckling, et al. Personality traits and brain health: a large prospective cohort study. *Nature Mental Health*, 1(10):722–735, 2023b.